Niveau(x) d'enseignement: 8 Numéro EDUQ : 6970202
Centre d'intérêt: 650

Ecole d'été en analyse des données exploratoire (1988 : Université du Québec à Montréal).

Actes de l'Ecole d'été sur l'analyse des données, juin 1988 / SCAD, CIRADE, [Ecole d'été en analyse des données exploratoire ; Edwin Diday, Yves Escoufier, Ludovic Lebart]. -- Montréal : Université du Québec à Montréal, Centre interdisciplinaire de recherches sur l'apprentissage et le développement en éducation, [1988?].
iii, 179 p. : ill. ; 29 cm.

Bibliogr. à la fin des textes.

Vedette(s)-secondaire(s): Diday, Edwin * Escoufier, Yves * Lebart, Ludovic * Université du Québec à Montréal. Service de consultation en analyse des données * Université du Québec à Montréal. Centre interdisciplinaire de recherches sur l'apprentissage et le développement en éducation

Descripteur(s): Analyse statistique * Analyse de données * Méthode de recherche * Sciences humaines * Sciences sociales * Autisme * Relations parents-école * Rôle de l'école * Education spéciale * Universitaire (Ordre d'enseignement)

Identificateur(s) de sujet: Analyse en composantes principales * Classification automatique * Analyse factorielle des correspondances

L'école d'été sur l'analyse des données tenue du 13 au 15 juin 1988 à l'UQAM s'adressait aux chercheurs, aux professeurs et aux étudiants diplômés oeuvrant plus particulièrement dans le domaine des sciences humaines et environnementales. Ces journées ont été consacrées aux méthodes d'analyse françaises plaçant l'accent sur la description des données. Ces actes présentent l'historique, les fondements et les domaines d'application de l'Analyse en composantes principales, la Classification automatique et l'Analyse factorielle des correspondances et comprend des exemples d'application à l'évaluation de l'acidité des lacs, à la géographie, à la linguistique, à l'étude de l'évolution des relations entre quatre familles et un service scolaire pour enfants autistiques de même qu'une étude comparative des trois approches.

Type de document: r

EDUQ
8970202

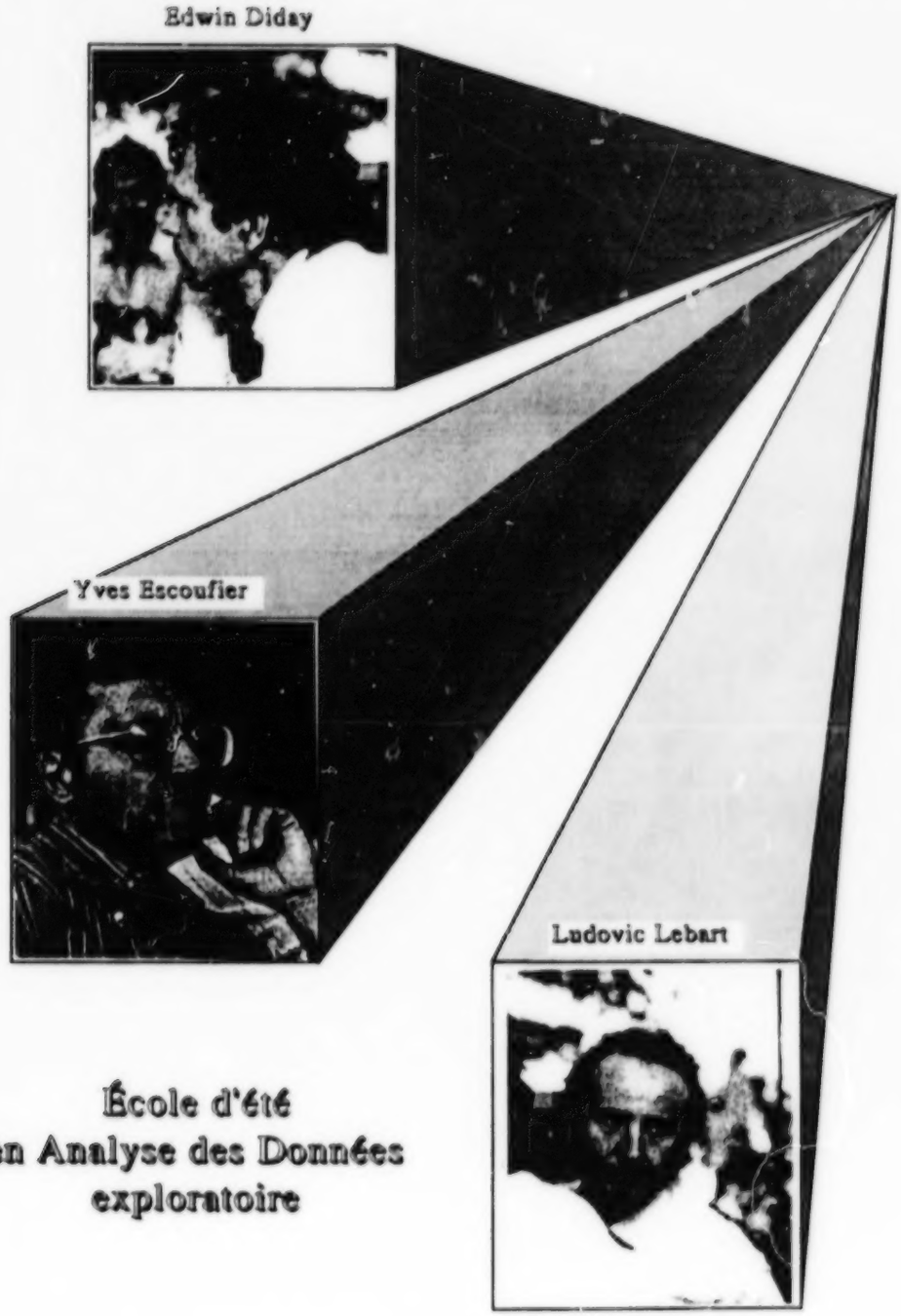
CENTRE INTERDISCIPLINAIRE DE RECHERCHES SUR L'APPRENTISSAGE ET LE DEVELOPPEMENT EN EDUCATION

ciradé

Université du Québec à Montréal

ACTES DE L'ÉCOLE D'ÉTÉ SUR L'ANALYSE DES DONNÉES, JUIN 1988,
SCAD-CIRADE

ACTES DE L'ÉCOLE D'ÉTÉ SUR L'ANALYSE DES DONNÉES, JUIN 1988, SCAD-CIRADE



École d'été
en Analyse des Données exploratoire

Table des matières

Table des matières 1

Introduction 1

Passé et futur de l'Analyse en Composantes Principales

Yves Escoufier 3

1 ACP d'ordre r de (Y, D) 3

1.1 Notations 3

1.2 Objectifs 3

1.3 4

2 ACP de (Y, Q, D) 5

2.1 5

2.2 5

3 Variables et individus supplémentaires 7

3.1 L'usage classique 7

3.2 Des usages moins classiques 7

4 Structure sur les individus 8

4.1 8

4.2 Les variables X sont quantitatives 8

4.3 Les variables X sont qualitatives 9

5 Structure sur les variables 10

5.1 Structure de contiguïté 10

6 Exemple d'application 12

6.1 Résultats de l'ACP sur les données centrées-réduites 12

6.2 Effet des essais 17

6.3 Élimination de l'effet taille 23

Références 27

La Classification Automatique

Edwin Diday 31

1 Les premières étapes d'une analyse de données 31

1.1 Les principaux choix de base 31

1.2 Construction d'un tableau de données 32

1.3 Notion d'inertie 34

1.4 Changement de variable et codage 37

1.5 Similarités et dissimilarités 39

2 Les méthodes de classifications automatiques 42

2.1 Introduction 42

2.2 Les principaux espaces de classification 43

2.3 La méthode des nuées dynamiques 45

2.4 La Méthode des Nuées Dynamiques dans le cas des centres de gravité 47

2.5 Classification avec distances adaptatives 50
2.6 La classification croisée 53
2.7 Classification d'un tableau de proximités 61
2.8 Interprétation d'une partition 64
2.9 Analyse d'une multi-partition 69
2.10 La classification hiérarchique 70
Conclusion 75
Bibliographie 76

L'analyse des correspondances
Ludovic Lebart 79
1 Principes de base des techniques d'a-nalyse des données 80
2 L'analyse des correspondances 80
2.1 Bref historique 80
2.2 Principes de base de l'Analyse des Correspondances exposés à partir d'un exemple numérique simple 81
2.3 Validité de la représentation 86
2.4 Variables actives et illustratives 88
3 Analyse des Correspondances Multiples 91
4 Complémentarité de la Classification 98
4.1 Les algorithmes de classification 99
4.2 Exemple d'application: Situations-types ou Noyaux factuels 102
Bibliographie sommaire 105

L'utilisation de l'analyse des correspondances multiples dans l'évaluation de l'acidification
Marius Lachance 107
1 Introduction 107
2 Données 108
3 Analyse de la variabilité spatiale de la qualité de l'eau 109
3.1 Choix des variables 109
3.2 Codage disjonctif des variables 110
3.3 Analyse des points variables 110
3.4 Analyse des points observations 112
3.5 Analyse avec une contrainte de voisinage 118
4 Recherche des facteurs explicatifs de la variabilité spatiale 119
4.1 Données considérées 120
4.2 Codage binaire des données 120
4.3 Calcul des tableaux de contingence 121
5 Élaboration d'un modèle de régression 126
5.1 Prévision des variables de qualité 128
5.2 Exemple d'application 129
6 Conclusion 130
Références 131

L'analyse exploratoire des données: de l'espace multidimensionnel à l'espace géographique
Guy Lemay 135
1 De la perception des espaces géographique et mathématique 135
2 Les difficultés géographiques: un poids deux mesures! 135
3 L'espace géographique multivarié 136
4 La démarche méthodologique intégrée 137
4.1 La matrice d'information spatiale: un atlas potentiel 137
4.2 La réduction structurelle 137
4.3 L'analyse typologique 138
4.4 La représentation visuelle 139
5 Au delà des espaces géographiques et mathématiques 139
Références 140

Étude des résiduelles de modèles log-linéaires par une généralisation de l'analyse des correspondances
Georges Hudon 145
1 Introduction 145
2 Les données 145
3 Modèles log-linéaires 145
4 Analyse des correspondances multiples 148
5 Analyse des correspondances multivariées 150
Bibliographie sommaire 154

Applications de l'analyse factorielle à la linguistique
Marc Bourdeau 155
1 Introduction 155
2 Analyse factorielle et phonologie 156
3 Phonétique en sociolinguistique 160
4 Conclusion 163
Références 164

Modélisation statistique d'une approche systémique: crise scolaire, métaphores verbales
Michelle Bourassa 167
1 Introduction: Pour une vision systémique d'une crise 167
2 Problématique 168
2.1 Caractéristiques fondamentales des systèmes souples 168
2.2 Caractéristiques fondamentales du système étudié 169
3 Analyse des correspondances 170
3.1 Choix épistémologiques 170
3.2 Type de données 171
3.3 Tableaux et graphiques 172
3.4 Analyse et discussion 174
4 Conclusion 180
Références 181

INTRODUCTION

Du 12 au 17 juin 1988, s'est tenue à l'Université du Québec à Montréal une école d'été sur l'Analyse de Données exploratoire. Cette école fut organisée conjointement par le Service de Consultation en Analyse de Données du département de mathématiques et informatique et le Centre Interdisciplinaire de Recherche sur l'Apprentissage et le Développement en Éducation. En plus de ressources fournies par le SCAD et le CIRADE, la tenue de cette école a été rendue possible grâce au soutien financier du Décanat des Études Avancés et de la Recherche, de la Fondation UQAM et du programme de la communauté scientifique réseau.

L'école s'adressait à l'ensemble des chercheurs, professeurs et étudiants gradués devant recueillir des données, les traiter et les analyser et plus particulièrement aux personnes oeuvrant dans le domaine des sciences humaines et environnementales.

Cette école aura permis d'initier les participants à la théorie et à la pratique des méthodes françaises d'analyse des données et à l'utilisation des logiciels SPAD, SICLA et LEAS. Les méthodes d'analyse française plaçant l'accent sur la description des données, ils disposeront désormais d'un outil statistique approprié pour explorer la structure des observations et formuler des hypothèses explicatives concernant les phénomènes étudiés.

Les journées du 13 au 15 juin ont été consacrées aux trois méthodes étudiées au cours de cette semaine: l'Analyse en Composantes Principales, la Classification Automatique et l'Analyse Factorielle des Correspondances. Durant la matinée de chacune de ces journées thématiques, un spécialiste oeuvrant dans le domaine de l'analyse des données a présenté les fondements d'une des méthodes, son historique et ses domaines d'application. Durant l'après-midi, les participants ont été invités à mettre en pratique à l'aide du logiciel SPAD les concepts exposés en les appliquant à un ensemble de données réelles. En plus des ateliers favorisant un contact direct avec le progiciel disponible sur l'ordinateur Amdahl, des démonstrations des logiciels existant sur micro-ordinateur ont eu lieu.

Messieurs YVES ESCOUFIER, EDWIN DIDAY et LUDOVIC LEBART présentèrent respectivement l'A.C.P, la C.A. et l'A.F.C. Pour leur part, le 16 juin, des chercheurs du réseau U.Q. utilisant ces méthodes (ou des méthodes apparentées) nous ont entretenu de cas réels qu'ils ont eu à traiter.

L'école d'été s'est terminée le 17 juin par une plénière où il était question de l'intérêt et des champs d'application de l'analyse exploratoire des données et par une conférence synthèse donnée par le professeur Escoufier.

Le comité organisateur de cette école tient à remercier l'ensemble des personnes ayant contribué, de près ou de loin, à la tenue ce cette semaine très intensive. Au risque d'en oublier quelques uns, il se doit de signaler le travail remarquable de

- PAULINE PROVENCHER-BANERJEE, MANON SÉNÉCAL, GUYLAINE ST-PIERRE et LISETTE GIROUX-ST-MARS du CIRADE,
- ANDRÉ OSTIGUY et LUC VAILLANCOURT du service informatique de l'UQAM,
- GUY CUCUMEL de l'Université Paris IX-Dauphine,
- GINETTE BÉGIN du SCAD

... et tous les participants!

Le comité organisateur,

- NADINE BEDNARZ, directrice du Cirade,
- JOCELYNE FERRARIS et ALAIN LATOUR, comité scientifique du SCAD.

PASSÉ ET FUTUR DE L'ANALYSE EN COMPOSANTES PRINCIPALES

Yves Escoufier,
Unité de Biométrie, Place Pierre Viala, 34060, Montpellier Cedex

Résumé

Après avoir rappelé les propriétés d'optimisation et d'approximation qui ont été à l'origine de l'Analyse en Composantes Principales dès 1901, on expose des approches récentes qui ouvrent des voies nouvelles dans le développement de cette technique: Structure sur les individus, Structure sur les variables, Structure de contiguïté.

1 ACP d'ordre r de (Y, D)

1.1 Notations

Considérons les données fournies en annexe 1 et arrêtons notre attention sur le tableau (n = 56) × (p = 6) des scores obtenus par les n individus sur les p variables.

Si **T** est ce tableau, nous appellerons **Y** le tableau déduit de **T** par centrage des colonnes. Si les individus sont munis de masse $m_i \geq 0$ ($\sum_{i=1}^{n} m_i = 1$) rangées dans la diagonale d'une matrice diagonale **D** on a:

$$\mathbf{Y} = (\mathbf{I}_{n\times n} - \mathbf{1}_n {}^t\mathbf{1}_n \mathbf{D})\mathbf{T}$$

où $\mathbf{I}_{n\times n}$ est la matrice identité d'ordre n, $\mathbf{1}_n$ le vecteur de $\mathbf{R}^n$ dont toutes les composantes sont normées à l'unité et ${}^t\mathbf{1}_n$ son transposé.

Avec ces notations la matrice des variances et covariances des variables est $\mathbf{V} = {}^t\mathbf{YDY}$. Nous noterons **W** la matrice $\mathbf{Y}{}^t\mathbf{Y}$ des produits scalaires entre les individus.

Par la suite $\mathbf{Y}_i$ représentera le vecteur ligne des p observations faites sur l'individu i et $\mathbf{Y}^j$ le vecteur colonne des mesures de la variable centrée occupant la $j^{ième}$ colonne.

1.2 Objectifs

On peut présenter l'ACP d'un tel tableau en fonction de plusieurs objectifs.

Objectifs d'optimisation

01: (K. PEARSON, (1901)). On veut trouver un sous espace $\mathbf{E}_r$ de $\mathbf{R}^p$ engendré par r vecteurs $v_1, \ldots, v_r$ de $\mathbf{R}^p$ de telle sorte que l'inertie de la projection du nuage sur un sous espace passant par l'origine et parallèle à $\mathbf{E}_r$ soit la plus grande possible. Notons $\mathbf{P}_r$ l'expression matricielle du projecteur sur $\mathbf{E}_r$ et $\mathcal{J}_r$ l'inertie de la projection du nuage dans $\mathbf{E}_r$. Le problème est donc:

- Trouver $\mathbf{E}_r$ qui maximise

$$\mathcal{J}_r = \sum_{i=1}^{n} m_i {}^t(\mathbf{P}_r {}^t\mathbf{Y}_i)(\mathbf{P}_r {}^t\mathbf{Y}_i).$$

02: (H. HOTELLING, (1933)). On veut trouver r variables combinaisons linéaires des variables initiales qui aient des variances les plus grandes possibles. Considérant r vecteurs orthonormaux v_α de $\mathbf{R}^p$ on peut donc traduire le problème sous la forme:

- Trouver r vecteurs $v_\alpha \in \mathbf{R}^p$, tels que
  $\forall (\alpha, \alpha')$: ${}^t v_\alpha v_{\alpha'} = \delta_{\alpha\alpha'}$ qui maximisent

$$\sum_{\alpha=1}^{r} \mathrm{Var}(\mathbf{Y}v_\alpha).$$

03: (OKAMOTO, (1969)). On veut trouver r variables orthogonales combinaisons linéaires des variables initiales qui aient des covariances

École d'été sur l'analyse des données, juin 1988, UQAM, SCAD-CIRADE.

avec les variables initiales les plus grandes possibles. On peut donc écrire:

- Trouver r vecteurs $v_\alpha \in \mathbb{R}^p$, tels que
  $\forall\ (\alpha, \alpha')$: ${}^t v_\alpha V v_{\alpha'} = 0$ et ${}^t v_\alpha v_\alpha = 1$ qui maximisent

$$\sum_{\alpha=1}^{r} \sum_{j=1}^{p} [\mathrm{Cov}(Y v_\alpha, X^j)]^2.$$

Objectifs d'approximations (R. Rao (1979))

A 1: On veut trouver $\tilde{V}_r$, matrice p × p de rang r qui minimise:

$$\sum_{j=1}^{p}\sum_{j'=1}^{p} (V_{jj'} - (\tilde{V}_r)_{jj'})^2$$
$$= \mathrm{Tr}[(V - \tilde{V}_r)(V - \tilde{V}_r)] = \|V - \tilde{V}_r\|^2.$$

A 2: On veut trouver $\tilde{W}_r$, matrice n × n de rang r qui minimise:

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} m_i m_{i'} (W_{ii'} - (\tilde{W}_r)_{ii'})^2$$
$$= \mathrm{Tr}[(W - \tilde{W}_r)D(W - \tilde{W}_r)D]$$
$$= \|(W - \tilde{W}_r)D\|^2$$

A 3: On veut trouver $\tilde{Y}_r$, matrice n × p de rang r qui minimise:

$$\sum_{i=1}^{n}\sum_{j=1}^{p} m_i (Y_{ij} - (\tilde{Y}_r)_{ij})^2$$
$$= \mathrm{Tr}[{}^t(Y - \tilde{Y}_r)D(Y - \tilde{Y}_r)]$$

1.3

Quel que soit l'objectif choisi pour introduire la méthode, les théorèmes généraux concernant les valeurs singulières d'une matrice (R. Sabatier (1987)) permettent de démontrer qu'on est toujours conduit à rechercher les mêmes r triplets ($\lambda_\alpha \in \mathbb{R}$, $\phi_\alpha \in \mathbb{R}^p$, $\psi^\alpha \in \mathbb{R}^n$) définis par:

- ${}^t YDY\phi_\alpha = \lambda_\alpha \phi_\alpha$ avec ${}^t\phi_\alpha \phi_{\alpha'} = \delta_{\alpha\alpha'}$
- $\psi^\alpha = \dfrac{Y\phi_\alpha}{\sqrt{\lambda_\alpha}}$
- $\lambda_1 \geq \lambda_2, \ldots, \geq \lambda_r$

On vérifie aisément que:

- $Y{}^t YD\psi^\alpha = \lambda_\alpha \psi^\alpha$ avec ${}^t\psi^\alpha D\psi^{\alpha'} = \delta_{\alpha\alpha'}$
- $\phi_\alpha = \dfrac{{}^t YD\psi^\alpha}{\sqrt{\lambda_\alpha}}$

Posons alors $\tilde{\phi}_\alpha = \sqrt{\lambda_\alpha}\phi_\alpha$, $\tilde{\psi}^\alpha = \sqrt{\lambda_\alpha}\,\psi^\alpha$ et $\Phi_{[r]} = (\phi_1 \vdots \ldots \vdots \phi_r)$.

0 1 est atteint pour $P_r = \Phi_{[r]}{}^t\Phi_{[r]}$ pour lequel:

$$\mathcal{J}_r = {}^t\Phi_{[r]} V \Phi_{[r]} = \sum_{\alpha=1}^{r} \lambda_\alpha$$

0 2 est atteint pour $\phi_1, \ldots, \phi_r$, $({}^t\phi_\alpha\phi_{\alpha'} = \delta_{\alpha\alpha'})$ et

$$\sum_{\alpha=1}^{r} \mathrm{Var}(Y\phi_\alpha) = \sum_{\alpha=1}^{r} \mathrm{Var}(\tilde{\psi}^\alpha) = \sum_{\alpha=1}^{r} \lambda_\alpha$$

0 3 est atteint pour les mêmes $\phi_1, \ldots, \phi_r$, $({}^t\phi_\alpha\phi_{\alpha'} = \delta_{\alpha\alpha'})$ et

$$\sum_{\alpha=1}^{r}\sum_{j=1}^{p} [\mathrm{Cov}(Y\phi_\alpha, Y^j)]^2$$
$$= \sum_{\alpha=1}^{r}\sum_{j=1}^{p} [\mathrm{Cov}(\psi^\alpha, Y^j)]^2 = \sum_{\alpha=1}^{r} \lambda_\alpha^2$$

A 1 est réalisé par

$$\tilde{V}_r = \sum_{\alpha=1}^{r} {}^t\tilde{\phi}_\alpha\tilde{\phi}_\alpha = \sum_{\alpha=1}^{r} \lambda_\alpha {}^t\phi_\alpha\phi_\alpha$$

et

$$\|V - \tilde{V}_r\|^2 = \sum_{\alpha=r+1}^{p} \lambda_\alpha^2.$$

A 2 est réalisé par

$$\tilde{W}_r = \sum_{\alpha=1}^{r} {}^t\tilde{\psi}^\alpha\,\tilde{\psi}^\alpha = \sum_{\alpha=1}^{r} \lambda_\alpha {}^t\psi^\alpha\psi^\alpha$$

et

$$\|(W - \tilde{W}_r)\,D\|^2 = \sum_{\alpha=r+1}^{p} \lambda_\alpha^2.$$

A 3 est réalisé par

$$\tilde{Y}_r = \sum_{\alpha=1}^{r} \sqrt{\lambda_\alpha}\,\psi^\alpha\,{}^t\phi_\alpha$$

et

$$\mathrm{Tr}[{}^t(Y - \tilde{Y}_r)D(Y - \tilde{Y}_r)] = \sum_{\alpha=r+1}^{p} \lambda_\alpha.$$

On doit remarquer dans tous ces énoncés la propriété d'emboîtement des résultats: la solution d'ordre r s'obtient en ajoutant le triplet (λ_r, ϕ_r, ψ^r) à la solution d'ordre r - 1.

Les $\tilde{\psi}^\alpha$ sont les <u>composantes principales</u>. Ce sont des variables de variance λ_α. Elles permettent l'approximation de W (objectif **A2**) et four-

nissent en conséquence les éléments pour la représentation des individus.

Les $\tilde{\phi}_\alpha$ sont les axes principaux. Ils peuvent être considérés comme des individus fictifs (R. RAO (1964)) reconstruisant la matrice V. Ils fourniront la représentation des variables.

Les objectifs **01** et **02** et l'égalité

$$\mathrm{Tr}(V) = \mathrm{Tr}(WD) = \sum_{\alpha=1}^{p} \lambda_\alpha$$

conduiront à dire qu'une valeur propre est une part de variance ou une part d'inertie expliquée par la composante principale.

2 ACP de (Y, Q, D)

2.1

Diverses raisons peuvent nous conduire à vouloir substituer à l'étude du tableau Y, celle d'un tableau $Z = YL$ déduit de Y par l'application d'une transformation linéaire L.

Quand par exemple les différentes variables ont des variances très différentes, l'étude des critères d'optimisation et d'approximation montre que les variables de grande variance prennent une part prépondérante dans l'analyse. On est alors conduit à préférer à Y, tableau des données centrées, le tableau $Z = Y[\mathrm{Diag}(V)]^{-1/2}$, tableau des données réduites.

Dans l'exemple qui nous intéresse nous pourrions vouloir travailler avec les accroissements $T_2 - T_1$, $T_3 - T_2$, $T_5 - T_4$, $T_6 - T_5$ plutôt qu'avec les scores eux-mêmes. Z serait alors le résultat de la multiplication de Y par la matrice

$$L = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

On est donc conduit à s'intéresser à l'ACP de $(Z = YL, D)$.

2.2

Posons $Q = L{}^tL$. On a

$$W_Z D = Z{}^tZD = YQ{}^tYD \text{ et } V_Z = {}^tLVL.$$

L'ACP d'ordre r de (Z, D) conduira au calcul des r triplets $(\lambda_\alpha, \eta_\alpha, \psi^\alpha)$ définis par:

${}^tLVL\eta_\alpha = \lambda_\alpha\eta_\alpha$ avec ${}^t\eta_\alpha\eta_{\alpha'} = \delta_{\alpha\alpha'}$

et

$W_Z D\psi^\alpha = \lambda_\alpha\psi^\alpha$ avec ${}^t\psi^\alpha D\psi^{\alpha'} = \delta_{\alpha\alpha'}$.

Si L est inversible on aura p triplets associés à des valeurs propres non nulles. Si L n'est pas inversible, on aura $q = \mathrm{rang}(L)$ valeurs propres non nulles.

On peut alors remarquer que si $VQ\phi_\alpha = \lambda_\alpha \phi_\alpha$ alors ${}^tLVL({}^tL\phi_\alpha) = \lambda_\alpha({}^tL\phi_\alpha)$.

C'est dire que $\eta_\alpha = {}^tL\phi_\alpha$ et qu'on aura ${}^t\eta_\alpha\eta_\alpha = 1$ si ${}^t\phi_\alpha Q\phi_\alpha = 1$.

Enfin de $\psi^\alpha = \dfrac{Z\eta_\alpha}{\sqrt{\lambda_\alpha}}$, il vient aisément que

$$\psi^\alpha = \frac{YQ\phi_\alpha}{\sqrt{\lambda_\alpha}}$$

d'où l'on tire:

$$\phi_\alpha = \frac{{}^tYD\psi^\alpha}{\sqrt{\lambda_\alpha}}.$$

Ces deux résultats constituent ce qu'on appelle les formules de transition.

Réécrivons alors tous les objectifs en fonction des ϕ_α en posant

$$N_{[r]} = (\eta_1 \mid \eta_2 \mid \ldots \mid \eta_r) = {}^tL\phi_{[r]}.$$

01: $\mathcal{J}_r = {}^tN_{[r]}V_Z N_{[r]} = {}^t\phi_{[r]}QVQ\phi_{[r]} = \sum_{\alpha=1}^{r} \lambda_\alpha$

02: est atteint par $\eta_1, \ldots, \eta_r$ $({}^t\eta_\alpha\eta_{\alpha'} = \delta_{\alpha\alpha'})$ donc par $\phi_1, \ldots, \phi_r$ $({}^t\phi_\alpha Q\phi_{\alpha'} = \delta_{\alpha\alpha'})$ et

$$\sum_{\alpha=1}^{r} \mathrm{Var}(Z\eta_\alpha) = \sum_{\alpha=1}^{r} \mathrm{Var}(Y\phi_\alpha) = \sum_{\alpha=1}^{r} \lambda_\alpha$$

03: est atteint par les mêmes vecteurs et

$$\sum_{\alpha=1}^{r}\sum_{j=1}^{p}[\mathrm{Cov}(Z\eta_\alpha, Y^j)]^2$$

$$= \sum_{\alpha=1}^{r}\sum_{j=1}^{p}[\mathrm{Cov}(Y\phi_\alpha, \tilde{Y}^j)]^2 = \sum_{\alpha=1}^{r}\lambda_\alpha^2$$

A1 est réalisé par

$$(\tilde{V}_Z)_r = \sum_{\alpha=1}^{r}\lambda_\alpha\eta_\alpha{}^t\eta_\alpha = \sum_{\alpha=1}^{r}\lambda_\alpha{}^tL\phi_\alpha{}^t\phi_\alpha L.$$

En posant $\tilde{V}_r = \sum_{\alpha=1}^{r}\lambda_\alpha\phi_\alpha{}^t\phi_\alpha$ on a:

$$\sum_{\alpha=r+1}^{q}\lambda_\alpha^2 = \mathrm{Tr}[(V_Z - (\tilde{V}_Z)_r)(V_Z - (\tilde{V}_Z)_r)]$$

$$= \mathrm{Tr}[(V - \tilde{V}_r)Q(V - \tilde{V}_r)Q]$$

$$= \|(V - \tilde{V}_r)Q\|^2$$

Pour **A2**, on a $\sum_{\alpha=r+1}^{q}\lambda_\alpha^2 = \|W_Z - (\tilde{W}_r)_r D\|^2$.

De même que D est incluse dans V et $\tilde{V}_r$ et n'apparait pas dans la notation, on peut décider de poser $W = YQ{}^tY$ et

$$\tilde{W}_r = \sum_{\alpha=1}^{r}\lambda_\alpha\psi^\alpha{}^t\psi^\alpha.$$

On aura alors:

$$\sum_{\alpha=r+1}^{p}\lambda_\alpha^2 = \|(W - \tilde{W}_r)D\|^2$$

Enfin pour **A3** posons:

$$\tilde{Z}_r = \sum_{\alpha=1}^{r}\sqrt{\lambda_\alpha}\psi^\alpha{}^t\eta_\alpha$$

$$= \sum_{\alpha=1}^{r}\sqrt{\lambda_\alpha}\psi^\alpha{}^t\phi_\alpha L = \tilde{Y}_r L$$

on a:

$$\sum_{\alpha=r+1}^{q}\lambda_\alpha = \mathrm{Tr}[{}^t(Z - \tilde{Z}_r)D(Z - \tilde{Z}_r)]$$

$$= \mathrm{Tr}[{}^t(Y - \tilde{Y}_r)D(Y - \tilde{Y}_r)Q]$$

Tout ceci conduit à parler de l'ACP de (Y, Q, D) qui aura pour composantes principales les vecteurs propres de $WD = YQ{}^tYD$ et pour axes principaux les vecteurs propres de $VQ = {}^tYDY\,Q$. Le paragraphe 1 traitait du cas où Q est l'identité de $\mathbb{R}^p$.

$\tilde{\psi}^\alpha = YQ\phi_\alpha$ fera qu'on sera conduit à parler des vecteurs $Q\phi_\alpha$ qui définissent la combinaison linéaire des variables à prendre en compte pour calculer les composantes principales: ces vecteurs sont appelés <u>facteurs principaux</u>.

$VQ\phi_\alpha = \lambda_\alpha\phi_\alpha$ avec ${}^t\phi_\alpha Q\phi_{\alpha'} = \delta_{\alpha\alpha'}$ implique

$$QV(Q\phi_\alpha) = \lambda_\alpha(Q\phi_\alpha)$$

avec $({}^t\phi_\alpha Q)Q^{-1}(Q\,\phi_{\alpha'}) = \delta_{\alpha\alpha'}$.

Ainsi les facteurs principaux sont vecteurs propres de QV et sont Q^{-1} orthogonaux.

Les formules:

$$\tilde{V}_r = \sum_{\alpha=1}^{r}\lambda_\alpha\phi_\alpha{}^t\phi_\alpha$$

$$\tilde{W}_r = \sum_{\alpha=1}^{r}\lambda_\alpha\psi^\alpha{}^t\psi^\alpha$$

et

$$\tilde{Y}_r = \sum_{\alpha=1}^{r}\sqrt{\lambda_\alpha}\,\psi^\alpha{}^t\phi_\alpha$$

sont <u>les formules de reconstitutions</u> de la matrice des variances et covariances, de la matrice des produits scalaires et des données. On remarquera que les formules d'approximations A1, A2, A3 font intervenir les matrices Q et D.

$$\lambda_\alpha = {}^t\tilde{\psi}^\alpha D\tilde{\psi}^\alpha = \sum_{i=1}^{n} m_i\,(\tilde{\psi}_i^\alpha)^2$$

conduit à définir

$$\frac{m_i(\tilde{\psi}_i^\alpha)^2}{\lambda_\alpha}$$

<u>contribution absolue de l'individu i</u> à la valeur propre λ_α. D'autre part,

$$W = \sum_{\alpha=1}^{p}\lambda_\alpha\psi^\alpha{}^t\psi^\alpha$$

implique en particulier que pour tout i:

$$W_{ii} = \sum_{\alpha=1}^{p} \lambda_\alpha (\psi_i^\alpha)^2 = \sum_{\alpha=1}^{p} (\tilde{\psi}_i^\alpha)^2$$

On appellera en conséquence

$$\frac{(\tilde{\psi}_i^\alpha)^2}{W_{ii}}$$

la contribution relative de la composante α à la reconstitution de W_{ii}, distance du point i au point moyen.

Symétriquement,

$$V = \sum_{\alpha=1}^{p} \lambda_\alpha \phi_\alpha {}^t\phi_\alpha$$

d'où $V_{jj} = \sum_{\alpha=1}^{p} (\tilde{\psi}_{\alpha j})^2$.

Aussi,

$$CTR = \frac{(\psi_{\alpha j})^2}{V_{jj}}$$

est la contribution relative de l'axe α à la reconstitution de la variance V_{jj} de la variable j.

On a de plus $\lambda_\alpha = {}^t\tilde{\phi}_\alpha Q \tilde{\phi}_\alpha$. Lorsque Q est diagonale on peut alors écrire

$$\lambda_\alpha = \sum_{j=1}^{p} Q_{jj} (\tilde{\psi}_{\alpha j})^2$$

et définir

$$CTA = \frac{Q_{jj}(\phi_{\alpha j})^2}{\lambda_\alpha}$$

contribution absolue de la variable j à la valeur propre λ_α.

3 Variables et individus supplémentaires

3.1 L'usage classique

Les variables sont représentées par les points ayant pour coordonnées les composantes des vecteurs $\tilde{\phi}_\alpha = {}^tYD\psi^\alpha$. Si la variable Y^j est réduite, $\tilde{\phi}_{\alpha j} = {}^tY^jD\psi^\alpha$ s'interprète comme le coefficient de corrélation linéaire entre Y^j et ψ^α. On utilise classiquement cette propriété pour donner un sens aux composantes principales (les interpréter).

Cette formule peut être étendue à une variable Y^{p+1} n'ayant pas participé à l'étude. Elle sera représentée par le point de coordonnées $\tilde{\phi}_{\alpha,p+1} = {}^tY^{p+1}D\psi^\alpha$.

Symétriquement la formule $\tilde{\psi}_\alpha = YQ\phi_\alpha$ permet de représenter les individus. Un individu supplémentaire Y_{n+1} n'ayant pas pris part à l'étude sera représenté par le point de coordonnées $\tilde{\psi}_{n+1}^\alpha = Y_{n+1}Q\phi_\alpha$.

Cette procédure est usuellement utilisée pour représenter les modalités d'une variable qualitative: on prend pour individu supplémentaire la moyenne des individus présentant la modalité.

3.2 Des usages moins classiques

3.2.1

GOWER et al. (1988) suggère de représenter la variable Y^j par l'individu supplémentaire ρe_j où $e_j \in \mathbb{R}^p$ a toutes ses composantes nulles exceptée la $j^{ième}$ qui est égale à 1. ρ est un scalaire qu'on peut faire varier. On a donc:

$$\tilde{\psi}_{(j)}^\alpha = \rho e_j Q \phi_\alpha = \rho Q_j \phi_\alpha = \rho (Q\phi_\alpha)_j$$

où Q_j et $(Q\phi_\alpha)_j$ sont respectivement les lignes j de Q et $Q\phi_\alpha$. Ceci revient à dire que la représentation de la variable sera donnée par la composante du facteur multipliée par ρ.

Si Q est l'identité et si on choisit $\rho = 1$ on retrouve la représentation traditionnelle.

Si Q est l'identité, l'utilisation répétée de cette procédure pour ρ variant par exemple de $\min(X^j)$ à $\max(X^j)$ donne des représentations très instructives.

Lorsque Q n'est pas l'identité, cette procédure permet de mettre en évidence l'importance prise par les variables initiales dans la définition de la composante principale. En effet Y^j peut être fortement corrélée à ψ^α ($\tilde{\phi}_{\alpha,j} = {}^tY^jD\psi^\alpha$ est grand) sans que Y^j contribue beaucoup à la variance de ψ^α ($(Q\phi_\alpha)_j$ est petit). On peut penser par

exemple à une matrice diagonale Q dont l'élément Q_{jj} est nul. Dans l'Analyse Discriminante, classiquement considérée comme l'ACP des points moyens des classes pour Q égale à l'inverse de la matrice des variances totales, cette procédure visualise clairement le rôle de chacune des variables dans la détermination des variables discriminantes.

3.2.2

Considérons K études $\{(\mathbf{Y}_k, \mathbf{Q}_k, \mathbf{D}) ; k = 1, \ldots, K\}$ concernant les mêmes individus munis des mêmes masses. Posons $\mathbf{W}_k\mathbf{D} = \mathbf{Y}_k\mathbf{Q}_k{}^t\mathbf{Y}_k\mathbf{D}$.

Soit $(\mathcal{Y}, \mathbf{Q}, \mathbf{D})$ l'étude définie par:

$$\mathcal{Y} = (\mathbf{Y}_1 \vdots \mathbf{Y}_2 \vdots \ldots \vdots \mathbf{Y}_K)$$

et

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 \ldots & 0 \\ \vdots & \ddots \vdots \\ 0 & \ldots \mathbf{Q}_k \end{pmatrix}$$

L'opérateur associé à cette étude est:

$$\mathcal{Y}\mathbf{Q}{}^t\mathcal{Y}\mathbf{D} = \sum_{k=1}^{K} \mathbf{W}_k\mathbf{D}$$

Posons $(\tilde{\mathcal{Y}}_k)_i = \frac{1}{K}(0 \vdots 0 \vdots (\mathbf{Y}_k)_i \vdots 0, \ldots, 0)$.

La coordonnée de l'individu i sur la composante ψ^α est alors:

$$\psi_i^\alpha = \mathcal{Y}_i\mathbf{Q}\phi_\alpha = \frac{1}{K}\sum_{k=1}^{K} (\tilde{\mathcal{Y}}_k)_i\mathbf{Q}\phi_\alpha$$

Ceci signifie que l'individu i apparait comme le point moyen d'individus fictifs qui seraient nuls (c'est à dire égaux à la moyenne) pour toutes les études excepté pour l'étude k.

On est donc en présence d'une version simplifiée de la méthode Statis (LAVIT (1988)) dans laquelle l'opérateur compromis est la somme des opérateurs associés à chaque étude.

4 Structure sur les individus

4.1

Reportons nous aux données fournies en annexe 2 et arrêtons notre attention sur le tableau (n = 56) × (q = 3) des variables concomitantes observées sur les individus. Ces variables ne décrivent pas les réponses des individus aux questions posées mais apportent une information sur les relations existant entre les individus eux-mêmes: deux individus présentent ou non la même modalité d'une variable qualitative, ils ont fourni ou non des valeurs voisines pour une variable quantitative. Notons que ces variables peuvent être envisagées dans deux contextes: ou bien on a tiré au hasard des individus dans une population et on a observé conjointement sur eux les variables définissant les colonnes de Y et X, ou bien, les variables définissant les colonnes de X constituent en fait un plan d'échantillonnage et c'est dans le cadre de ce plan que les variables Y ont été observées.

On peut légitimement penser que l'observation conjointe du tableau Y et de ce tableau X est motivée par le désir d'étudier dans quelle mesure la connaissance de X détermine celle de Y.

La démarche classique évoquée en 3 serait d'utiliser ces variables en éléments supplémentaires dans l'ACP du tableau Y. On peut juger plus rigoureux de prendre en compte ces informations dès le début de l'étude.

4.2 Les variables X sont quantitatives

LEBART et al. (1971) traite de l'Analyse des covariances partielles. On peut aborder le problème comme un problème d'approximation en se demandant quelle est la part de l'ACP de Y qui pourrait être reconstruite à partir de X. Nous avons vu que dans l'ACP de Y, la représentation des individus était déterminée par les vecteurs propres de l'opérateur $\mathbf{WD} = \mathbf{YQ}{}^t\mathbf{YD}$. On est donc conduit à chercher la métrique R à associer

à X de telle sorte que $\|YQ^tYD - XR^tXD\|^2$ soit minimum.

Une démonstration simple dont on trouvera les détails en BONIFAS et al. (1984) établit que:

$$R = ({}^tXDX)^{-1}\,{}^tXDWDX({}^tXDX)^{-1}$$

Ceci montre, entre autres choses, que du point de vue de la représentation des individus l'étude (X, R, D) est équivalente à l'étude (P_XY, Q, D) où $P_X = X({}^tXDX)^{-1}\,{}^tXD$ est le projecteur D-orthogonal sur le sous espace de $\mathbf{R}^n$ engendré par les colonnes de X. En effet les deux études conduisent au même opérateur P_XWDP_X.

Ayant étudié la part de l'ACP de Y qui peut être reconstruite à partir de X, on peut vouloir étudier la part de cette ACP qui n'a pas été reconstruite par X, c'est à dire qui se trouve dans le sous-espace de $\mathbf{R}^n$ orthogonal au sous-espace engendré par les colonnes de X. On considère donc l'ACP de ($(I-P_X)Y$, Q, D) d'opérateur associé $(I-P_X)WD(I-P_X)$.

Posons $V = {}^tYDY$, $V_{P_X} = {}^t(P_XY)D(P_XY)$ et $V_{(I-P_X)} = {}^t((I-P_X)Y)D((I-P_X)Y)$. On a de façon évidente $V = V_{P_X} + V_{(I-P_X)}$ ce qui entraîne en particulier que les études (P_XY, Q, D) et ($(I-P_X)Y$, Q, D) décomposent l'inertie de l'étude (Y, Q, D).

Les composantes principales de (P_XY, Q, D) sont des combinaisons linéaires des projections des variables de Y sur le sous-espace engendré par X: elles appartiennent donc à ce sous-espace. De même les composantes principales de ($(I-P_X)Y$, Q, D) appartiennent au sous-espace de $\mathbf{R}^n$ orthogonal au sous espace engendré par X.

En ce sens ces deux études sont orthogonales.

Dans le cas particulier où $Q = V^{-1}$:

$$V_{P_X}V^{-1}\phi = \lambda\phi \Leftrightarrow (V - V_{(I-P_X)})V^{-1}\phi = \lambda\phi$$
$$\Leftrightarrow V_{(I-P_X)}V^{-1}\phi = (1 - \lambda)\phi$$

Aux plus grandes valeurs propres de l'une des études correspondent donc les plus petites valeurs propres de l'autre.

Dans le cas particulier où $Q = I_{p\times p}$, la variance de chacune des variables initiales est décomposée en deux parts selon l'équation

$$V = V_{P_X} + V_{(I - P_X)}.$$

Chacune de ces variances partielles est elle-même décomposée sur les composantes principales des études (P_XY, $I_{p\times p}$, D) et ($(I-P_X)Y$, $I_{p\times p}$, D).

4.3 Les variables X sont qualitatives

4.3.1

Soit U, $n \times q$, le tableau des variables indicatrices associé à une variable qualitative à q modalités. Par définition chaque ligne de U comporte un 1 dans la colonne correspondant à la modalité prise par l'individu et des 0 ailleurs. C'est dire en particulier que 1_n appartient au sous-espace [U] de $\mathbf{R}^n$ engendré par les colonnes de U. Un résultat classique donne alors:

$$1_n \subset [U] \Rightarrow P_{1_n}P_U = P_UP_{1_n} = P_{1_n}$$

et

$$P_{U\cap 1_n^{\perp}} = P_U - P_{1_n}$$

Comme en 4.2, on peut vouloir étudier l'ACP des études (P_UY, Q, D) et ($(I-P_U)$ Y, Q, D). Les résultats prennent une signification particulière liée à la nature des tableaux P_UY et $(I-P_U)Y$.

En effet,

$$P_UY = U({}^tUDU)^{-1}\,{}^tUDY$$
$$= U{}^tM$$

où M est la matrice $q \times p$ des moyennes des variables Y dans chacune des modalités de la variable qualitative. P_UY n'est d'autre qu'un tableau dans lequel chaque individu est décrit par le point moyen de la modalité qu'il présente.

$(I-P_U)Y$ est le tableau des écarts à ces valeurs moyennes.

Les deux études décomposent l'inertie de (Y, Q, D). Le choix particulier $Q = V^{-1}$ correspond à la méthode d'analyse discriminante.

4.3.2

Considérons maintenant le cas où deux variables qualitatives U_1 et U_2 ont été observées sur les individus.

On notera $[U_{12}]$ le sous-espace de $\mathbf{R}^n$ engendré par les colonnes de U_1 et U_2 et on a bien sûr:

$$\mathbf{R}^n = [U_{12}] \oplus [U_{12}]^{\perp} .$$

Une étude qui veut permettre de retrouver dans l'ACP de (Y, Q, D) les parts expliquées par U_1, U_2 et leurs interactions, conduit à chercher des décompositions en sommes directes de $[U_{12}]$. On retrouve là des problématiques usuelles de l'Analyse de Variance Multiple (SCHEFFÉ, (1959)). Les deux facteurs pourront être croisés ou hiérarchisés. Des facteurs croisés pourront être orthogonaux ou non. On pourra se préoccuper d'interactions ou non.

Dans tous les cas, la difficulté est d'expliciter une décomposition de $[U_{12}]$ en somme directe de sous-espaces puis d'écrire les projecteurs sur ces sous-espaces.

On trouvera une étude théorique de ces points de vue et une application détaillée dans R. SABATIER, (1987). CHESSEL D. (1987) et LEBRETON J.D. (1988) donnent des applications originales de ces résultats.

5 Structure sur les variables

Nous terminerons cette présentation en évoquant le cas d'une structure donnée sur les variables. On peut imaginer par exemple dans un contexte d'enquête sociologique des variables résultant d'un questionnaire organisé en chapitres homogènes. Dans une application pédologique les variables peuvent être liées par la profondeur à laquelle elles ont été observées, par la procédure d'observation et éventuellement même par l'observateur.

Tant que ces observations peuvent être décrites par la donnée d'une variable quantitative ou qualitative sur $\mathbf{R}^p$ leur prise en compte est permise par l'introduction des projecteurs Π et $\Pi^{\perp}$ qui leur sont associés. L'ACP de $(Y\Pi, Q, D)$ fournira une représentation des individus bâtis sur leur projection dans le sous espace de $\mathbf{R}^p$ engendré par la structure considérée.

L'ACP de $(Y\Pi^{\perp}, Q, D)$ permet d'étudier ce qui dans la description des individus n'est pas lié à la structure. Une application très simple de cette approche est l'élimination d'un effet taille dans une analyse. Prenant $\Pi = \frac{1}{p} 1_p {}^t 1_p$, $Y\Pi$ est le tableau $n \times p$ dans lequel chaque individu est décrit par p composantes égales à la moyenne des valeurs qu'il a prise.

$Y\Pi^{\perp}$ est alors le tableau des données centrées par lignes. On voit que de telles démarches consistant à introduire un projecteur dans $\mathbf{R}^p$ peuvent être combinées avec l'introduction de projecteur dans $\mathbf{R}^n$. On sera alors conduit de manière générale à l'ACP du triplet $(PY\Pi, Q, D)$ le choix de P et Π étant guidé par les motivations de l'étude.

5.1 Structure de contiguïté

Certaines applications conduisent à considérer des individus sur lesquels sont définis des liens naturels pouvant être décrits par un graphe. Les individus pourront être par exemple les noeuds d'un réseau routier; les variables sont les nombres de véhicules comptés en ces noeuds à différentes heures de la journée ou en différentes occasions; les liens sont les routes qui relient effectivement les noeuds (A. MOM, (1988)). Dans d'autres applications les individus seront des zones géographiques qui seront considérées comme liées lorsqu'elles sont mitoyennes. On

trouvera dans L. LEBART (1969) la première introduction de cette problématique en ACP.

Suivant A. MOM, définissons $\alpha_{(i)}$ l'ensemble des voisins de l'individu i sur le graphe noté G_1 et $|\alpha_{(i)}|$ le nombre d'éléments de $\alpha_{(i)}$. On pose $K_1 = \sum_{i=1}^{n} |\alpha_{(i)}|$. Le graphe étant ici considéré comme non orienté, K_1 est le double du nombre d'arêtes de G_1.

L_1 sera la matrice $n \times n$ d'incidence du graphe; $(L_1)_{ii'}$ vaut 1 si i et i' sont reliés et 0 dans le cas contraire. Pour tout i, $(L_1)_{ii} = 0$. L_1 est symétrique puisque le graphe est non orienté.

Soit N_1 la matrice $n \times n$ diagonale d'élément $(N_1)_{ii} = |\alpha_{(i)}|$, on peut alors définir une matrice V_1, de dimension $p \times p$ par:

$$V_1 = {}^t Y \frac{N_1 - L_1}{K_1} Y$$

on a

$$(V_1)_{jj'} = \frac{1}{2K_1} \sum_{i=1}^{n} \sum_{k \in \alpha_{(i)}} (Y_i^j - Y_k^j)(Y_i^{j'} - Y_k^{j'})$$

Ceci montre que V_1 s'interprète comme une matrice de variances et covariances liée au graphe G_1. En particulier:

$$(V_1)_{jj} = \frac{1}{2K_1} \sum_{i=1}^{n} \sum_{k \in \alpha_{(i)}} (Y_i^j - Y_k^j)^2 \geq 0$$

Au graphe des contiguïtés G_1, on peut associer le graphe des non contiguïtés G_2. On notera N_2 et L_2 les matrices analogues à N_1 et L_1 qui lui sont associées et K_2 l'analogue de K_1.

On définit:

$$N = N_1 + N_2 = (n-1)\, I_{n \times n}$$
$$L = L_1 + L_2 = 1_n {}^t 1_n - I_{n \times n}$$
$$K = K_1 + K_2 = n(n-1)$$

En posant $V_2 = {}^t Y \frac{N_2 - L_2}{K_2} Y$, on a alors:

$$\frac{n-1}{n}\left(\frac{K_1 V_1 + K_2 V_2}{K}\right)$$
$$= \frac{K_1 V_1 + K_2 V_2}{n^2}$$
$$= {}^t Y \frac{(N_1 + N_2 - L_1 + L_2)}{n^2} Y$$
$$= {}^t Y \frac{(n 1_n - \frac{1}{n} 1_n {}^t 1_n)}{n^2} Y$$
$$= \frac{1}{n} {}^t Y (I_n - \frac{1}{n} 1_n {}^t 1_n) Y = V.$$

On dira par la suite que V_1 est la matrice de variances et covariances locales ; V_2 est la matrice résiduelle ; V est la matrice totale.

La semi-définie positivité de $(N_1 - L_1)/K_1$ permet d'affirmer qu'il existe une matrice M, de dimension $\text{rang}(V_1) \times p$ telle que

$$\frac{N_1 - L_1}{K_1} = \frac{{}^t MM}{n}.$$

Ceci conduit à envisager des ACP fondées sur le tableau MY dans lesquelles les individus ont tous le même poids $\frac{1}{n}$. La matrice de variance associée sera V_1 et ne prendra donc en compte que les variances et covariances locales.

En particulier, les premiers facteurs principaux de l'ACP de

$$(MY, V^{-1}, \tfrac{1}{n} I_{n \times n})$$

que nous noterons u_α, sont $(V^{-1})^{-1} = V$ orthonormaux: ${}^t u_\alpha V u_{\alpha'} = \delta_{\alpha\alpha'}$ c'est dire qu'ils définissent des combinaisons linéaires des variables Yu_α qui sont orthonormales. Ils définissent par ailleurs les composantes principales $\tilde{\psi}_\alpha = MYu_\alpha$ qui sont telles que:

$$\frac{{}^t \tilde{\psi}_\alpha \tilde{\psi}_{\alpha'}}{n} = {}^t u_\alpha {}^t Y \frac{{}^t MM}{n} Y u_{\alpha'}$$
$$= {}^t u_\alpha V_1 u_{\alpha'} = \delta_{\alpha\alpha'} \lambda_\alpha$$

Ces résultats peuvent être exprimés en disant que les facteurs principaux définissent des combinaisons linéaires qui sont de variance locale maximale sous la contrainte d'être de variance (totale) égale à 1. La relation

$$\frac{K_1V_1 + K_2V_2}{n^2} = V$$

implique de plus que ces combinaisons linéaires sont de variances résiduelles minimales.

6 Exemple d'application

Dans cette section nous présentons les résultats des analyses produites par le logiciel LEAS sur un fichier de données 56×6, illustrant les méthodes introduites dans le texte. Les données brutes sont présentées aux annexes 1 et 2.

Les individus concernent 56 personnes soumises à un test psychologique pour mesurer leur état dépressif avant et après traitement (3 traitements thérapeutiques différents). Les 3 premières variables correspondent aux scores obtenus à une question précise posée durant trois périodes successives et les 3 dernières variables correspondent au total des scores obtenus à l'ensemble des questions au cours des mêmes périodes.

6.1 Résultats de l'ACP sur les données centrées-réduites

Dans un premier temps nous présentons les résultats obtenus par l'ACP avec les données centrées et réduites.

Tableau 6.1: Données centrées réduites

T_1	T_2	T_3	T_4	T_5	T_6	T_1	T_2	T_3	T_4	T_5	T_6
0.151	-0.425	-0.362	1.443	-0.138	-1.116	-0.791	-0.022	-1.098	-0.507	0.480	-0.197
0.151	0.381	0.743	0.577	0.788	0.460	0.151	0.784	-1.466	-0.940	0.171	-1.642
0.151	-0.425	-0.362	0.360	-0.292	-0.328	0.623	0.784	1.480	-0.290	0.325	1.773
2.036	2.799	2.216	2.960	3.103	1.773	0.151	-0.022	-0.362	-0.074	-0.292	-0.197
0.151	0.381	-0.362	0.793	0.480	-0.328	0.151	0.784	1.111	-0.074	0.171	0.722
0.151	0.381	1.111	0.793	0.325	0.985	-1.733	-0.425	-0.362	-1.157	-0.292	0.591
0.623	1.993	1.480	0.793	0.788	1.116	-0.791	-1.634	-0.362	0.143	-1.681	-0.328
-1.262	-0.425	-1.098	-0.940	-0.755	-0.722	0.151	0.381	0.375	0.360	0.942	1.248
2.036	1.993	1.848	1.443	1.714	1.773	-0.320	-1.231	-1.466	-0.507	-1.527	-1.379
1.565	0.784	-0.362	1.010	0.480	-0.854	-2.205	-1.634	-1.466	-0.940	-1.527	-1.379
0.151	-0.022	0.375	0.143	-0.446	-0.197	-0.791	-1.231	-1.098	-0.507	0.171	0.328
0.151	1.187	1.480	-0.074	1.097	1.510	0.151	-0.828	0.375	-0.940	-1.681	-0.591
1.565	1.590	1.848	0.577	0.788	1.116	0.151	0.381	0.743	-0.507	0.634	0.328
-0.320	0.381	0.743	0.577	0.788	0.985	0.151	-1.231	-1.466	-0.507	-1.527	-1.510
-1.262	-0.425	0.007	-0.074	0.325	0.591	-0.320	-0.828	-1.098	-1.157	-0.909	-0.985
2.507	1.590	-0.730	3.177	1.868	-0.854	-0.791	0.784	-1.098	-0.290	0.942	-1.510
2.036	1.590	1.111	2.093	1.714	0.460	0.623	-1.231	1.480	-0.724	-1.218	0.197
0.151	-0.022	-0.362	0.360	0.480	-0.854	-2.676	0.381	-0.730	-2.241	-0.292	-0.985
0.623	0.381	-1.466	1.660	0.480	-1.379	-0.791	-1.231	-0.362	-0.724	-1.064	-0.591
1.094	1.187	2.216	-0.074	0.325	2.693	0.151	-0.425	-0.730	-1.157	-1.527	-1.248
-0.320	-2.036	0.007	-0.074	-1.681	-0.066	0.623	-0.425	-1.098	-1.157	-0.755	-0.985
0.151	-0.425	0.007	0.360	-0.292	0.722	-0.320	0.381	0.007	-1.157	-0.601	-0.197
0.151	-0.425	0.743	-0.074	-0.138	0.591	0.151	-0.828	-1.098	-0.940	-0.909	-1.248
-0.320	-0.022	0.007	-0.290	0.480	0.722	0.151	-1.231	-0.730	-0.074	-1.064	-0.854
-0.791	-0.828	0.375	-0.507	-0.755	0.460	0.151	-0.425	0.375	-0.724	-0.138	-0.197
0.623	-0.425	0.743	0.360	-0.601	1.116	-1.262	-0.828	-0.362	-0.940	-0.446	-0.460
0.623	0.381	0.007	0.793	-0.138	-0.591	-1.262	-0.828	-0.362	-0.290	0.788	0.460
-0.791	-0.022	-0.362	0.143	1.097	0.722	-1.262	0.784	-0.730	-0.290	0.942	0.328

Tableau 6.2: Caractéristiques des variables

Variable	Moyenne	Variance	Écart-type	Minimum	Maximum	Données manquantes
T_1	0.000	1.000	1.000	-2.676	2.507	0 ind.
T_2	0.000	1.000	1.000	-2.036	2.799	0 ind.
T_3	0.000	1.000	1.000	-1.466	2.216	0 ind.
T_4	0.000	1.000	1.000	-2.241	3.177	0 ind.
T_5	0.000	1.000	1.000	-1.681	3.103	0 ind.
T_6	0.000	1.000	1.000	-1.642	2.693	0 ind.

Tableau 6.3: Matrice des corrélations

	T_1	T_2	T_3	T_4	T_5	T_6
T_1	1.000					
T_2	0.580	1.000				
T_3	0.498	0.546	1.000			
T_4	0.719	0.574	0.371	1.000		
T_5	0.436	0.848	0.437	0.662	1.000	
T_6	0.271	0.466	0.851	0.288	0.509	1.000

Tableau 6.4: Édition des premières valeurs propres

Somme des valeurs propres: 6.00000
Somme des carrés des val. pr.: 15.5222

Val. prop.	Inertie(%)	% cumulés
3.70002	61.67	61.67
1.11630	18.61	80.27
0.67483	11.25	91.52
0.33854	5.64	97.16
0.11120	1.85	99.01
0.05911	0.99	100.00

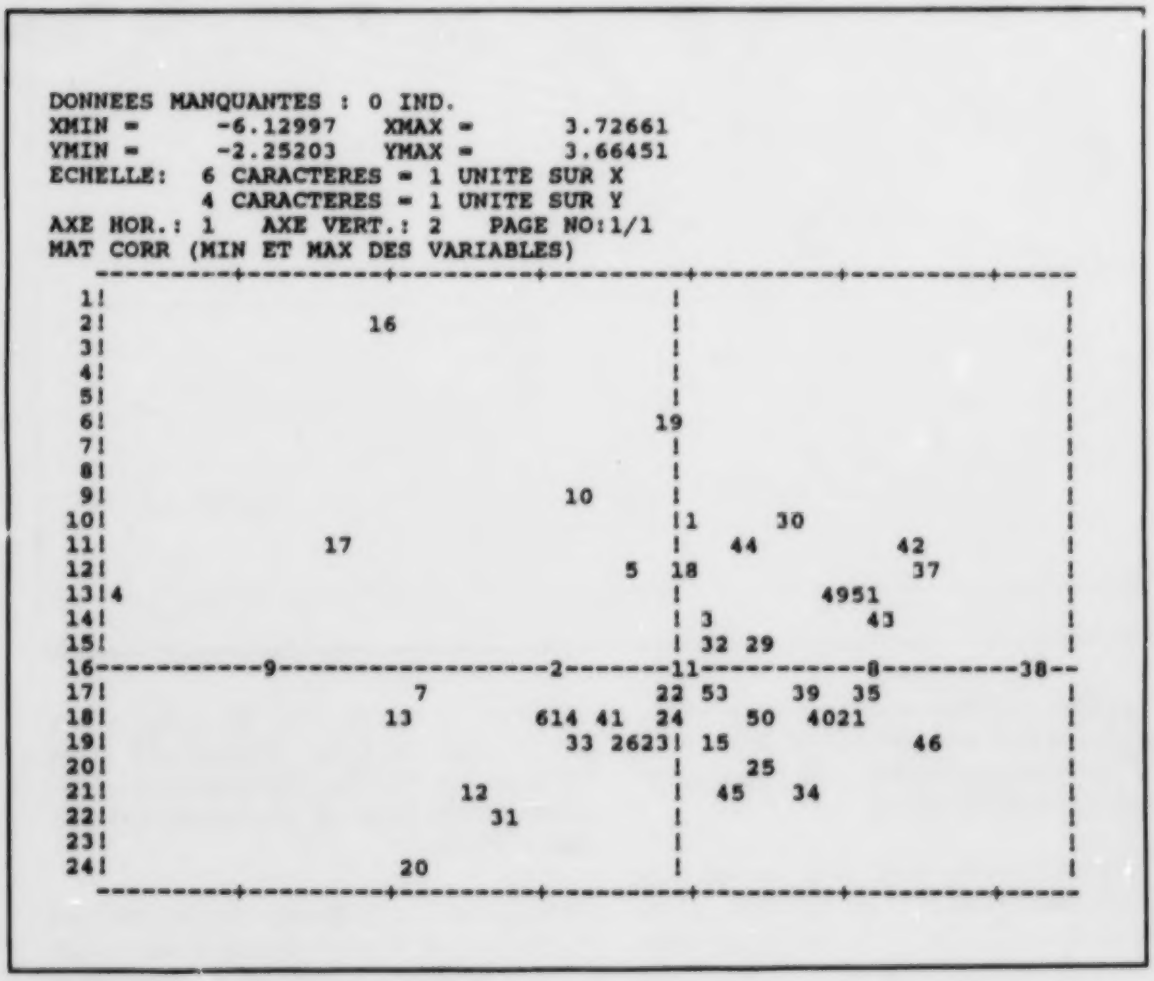
Tableau 6.5: Description des variables actives

	Axe 1	CTA	CTR	Axe 2	CTA	CTR	Axe 3	CTA	CTR
T_1	-0.74	148	548	0.36	119	133	0.50	374	252
T_2	-0.87	202	749	0.14	16	18	-0.32	150	101
T_3	-0.78	163	603	-0.53	254	284	0.26	97	65
T_4	-0.77	160	592	0.47	200	223	0.15	32	22
T_5	-0.84	190	705	0.16	22	24	-0.48	347	234
T_6	-0.71	136	503	-0.66	389	434	-0.00	0	0
Moyenne			616.7			186.1			112.5
Moyenne cumulée			616.7			802.7			915.2

Tableau 6.6: Description des individus actifs
(CP ↔ Composantes principales)

	CP1	CTA	CTR	CP2	CTA	CTR	CP3	CTA	CTR
1	0.17	0	4	1.80	36	613	0.49	6	65
2	-1.27	9	661	-0.19	1	10	-0.19	1	19
3	0.38	1	224	0.80	4	385	0.38	4	223
4	-0.13	101	978	0.62	4	10	-0.45	5	5
5	-0.49	1	189	0.91	13	659	-0.31	2	74
6	-1.80	15	725	-0.67	7	145	0.24	1	10
7	-2.81	38	871	-0.80	4	28	-0.26	2	7
8	2.09	21	803	-0.02	0	0	-0.67	12	91
9	-4.41	94	985	-0.18	1	2	0.29	2	4
10	-1.11	6	237	1.88	66	681	0.64	5	38
11	0.01	0	0	-0.02	0	1	6.81	7	609
12	-2.20	23	678	-1.35	29	288	-(.57	9	46
13	-3.08	45	878	-0.51	4	25	0.55	8	29
14	-1.29	8	607	-0.68	7	160	-0.48	6	83
15	0.34	1	53	-0.69	11	321	-0.81	17	296
16	-3.16	48	421	3.66	215	568	0.16	1	1
17	-2.70	44	888	1.25	25	101	0.34	3	7
18	0.06	0	3	1.00	16	798	-0.23	1	41
19	-0.18	0	4	2.68	115	946	-0.20	1	5
20	-2.96	42	587	-2.25	81	341	0.68	12	31
21	1.82	16	470	-0.61	6	53	1.57	65	349
22	-0.15	0	25	-0.34	3	122	0.49	4	289
23	-0.30	0	77	-0.80	10	569	0.55	8	272
24	-0.23	0	56	-0.63	6	417	-0.82	7	391
25	0.89	4	317	-1.19	23	971	0.31	2	58
26	-0.64	2	144	-0.84	11	346	1.19	38	496
27	-0.45	1	134	0.96	15	606	0.46	4	139
28	-0.34	1	47	-0.32	3	46	-1.21	39	886
29	0.82	3	288	0.25	1	26	-1.19	37	690
30	1.09	6	185	1.82	37	361	-0.93	23	139
31	-1.87	17	536	-1.62	42	481	0.29	2	13
32	0.33	1	377	0.28	1	274	0.15	1	77
33	-1.17	7	565	-0.87	12	389	0.02	0	0
34	1.38	9	372	-1.40	31	384	-1.05	29	215
35	1.98	19	614	-0.28	1	12	1.05	29	174
36	-1.40	9	674	-0.87	5	111	-0.43	5	65
37	2.65	34	848	0.88	12	94	0.64	11	49
38	3.73	67	939	-0.02	0	0	-0.44	5	13
39	1.31	9	458	-0.28	1	21	-0.54	8	79
40	1.49	11	483	-0.54	5	60	1.38	48	373
41	-0.73	3	383	-0.61	6	352	-0.29	2	86
42	3.51	38	738	1.13	20	148	0.93	23	100
43	3.16	23	911	0.30	1	18	0.12	0	3
44	0.66	2	76	1.33	26	311	-1.73	79	885
45	0.48	1	35	-1.32	38	282	1.90	96	891
46	2.54	31	464	-0.94	14	63	-2.24	132	360
47	1.98	19	913	-0.36	2	38	0.28	4	34
48	2.02	20	682	0.40	3	27	0.73	14	89
49	1.55	12	516	0.70	8	105	0.44	5	43
50	0.75	3	280	-0.55	5	151	-0.19	1	19
51	1.99	19	764	0.72	8	101	0.44	5	38
52	1.40	12	669	0.61	6	93	0.96	24	234
53	0.48	1	176	-0.41	3	184	0.33	3	115
54	1.74	15	882	-0.96	5	84	-0.47	4	59
55	0.41	2	111	-0.66	7	131	-1.08	31	352
56	0.01	0	0	-0.18	0	7	-1.91	97	956
MOYENNE			449.3			227.2			169.4
MOYENNE CUMULEE			449.2			676.4			845.9

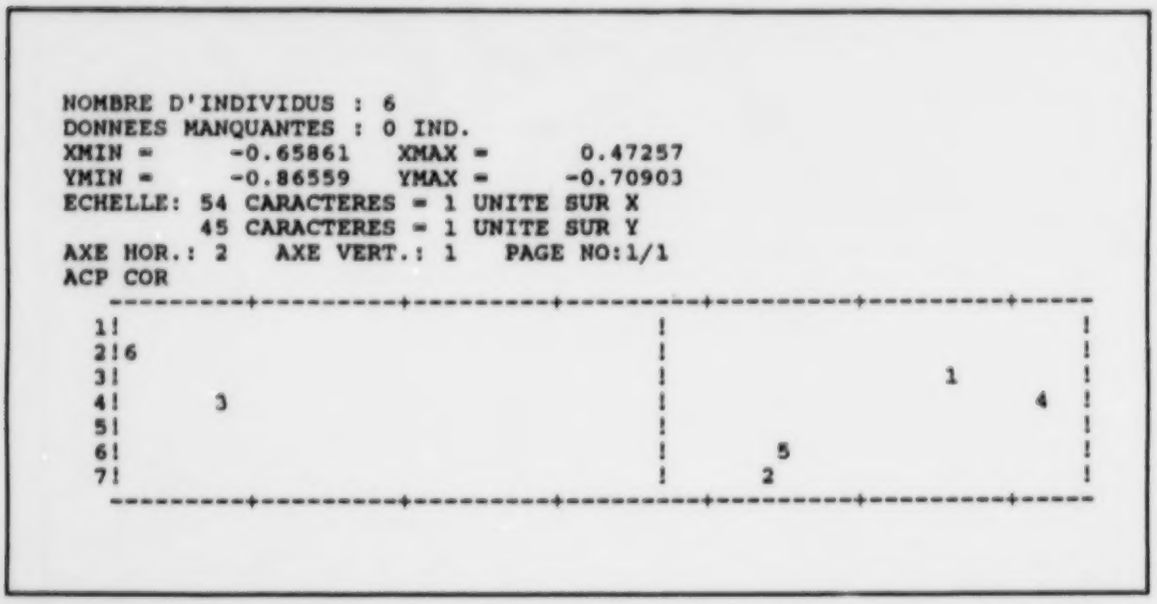
Figure 6.1: Représentation des 56 individus actifs sur les deux premiers axes



LISTE DES POINTS CACHES

IDENTIF.	LIGNE	COLONNE	PAGE
27	12	36	1
28	17	37	1
36	18	30	1
47	17	51	1
48	14	51	1
52	13	49	1
54	18	50	1
55	18	43	1
56	16	39	1

Figure 6.2: Représentation des variables actives sur les deux premiers axes



Les deux premières valeurs propres représentant plus de 80% de l'inertie totale, nous limiterons ici notre commentaire aux composantes principales correspondantes.

Le premier facteur principal $\widetilde{\phi}_1$ a toutes ses composantes négatives et pratiquement égales. Ceci conduit à dire que la composante principale $\widetilde{\psi}^1$ sera une composante d'intensité globale des scores (on dit une composante de taille). Les individus qui auront eu toujours des scores supérieurs à la moyenne donneront des valeurs négatives pour $\widetilde{\psi}^1$. Ceux qui auront eu toujours des scores inférieurs à la moyenne donneront des valeurs positives pour $\widetilde{\psi}^1$.

Le facteur $\widetilde{\phi}_2$ oppose principalement les variables 1 et 4 aux variables 3 et 6. $\widetilde{\psi}^2$ sera donc positif pour des individus ayant des scores supérieurs à la moyenne dans la première période et inférieurs à la moyenne dans la troisième (leur état s'est amélioré au cours du temps). $\widetilde{\psi}^2$ négatif traduit une évolution défavorable des individus: des scores faibles dans la période 1 et forts dans la période 3.

La figure 6.1 attire l'attention sur les individus extrêmes 4, 16, 38 et 20. On retournera aux fichiers de données pour vérifier qu'ils supportent bien cette interprétation.

La figure 6.2 est celle des représentations des variables faites selon la procédure décrite en 3.2.1 en prenant pour chaque variable les points correspondants à son minimum et son maximum. On voit que les segments associés à chaque variable ont sensiblement la même longueur ce qui montre que les variables ont des rôles comparables dans l'étude.

6.2 Effet des essais

Pour étudier l'effet des essais sur les résultats calculons d'abord les moyennes des 6 variables dans chacun des essais (cf. Tableau 6.7).

La première procédure (cf. 3.1) consiste à projeter ces points en supplémentaires dans l'ACP précédente. Les essais 1, 2 et 3 sont respectivement représentés par les points 57, 58 et 59 (figure 6.3).

On voit que la moyenne de l'essai 1 est plutôt vers les scores forts tandis que 59 est plutôt dans les scores faibles. On pourrait bien sûr utiliser ici des procédures de tests paramétriques, non paramétriques ou des méthodes de rééchantillonnage pour apprécier l'intensité de ces différences. Nous nous contenterons ici d'une approche plus exploratoire telle qu'elle est décrite en 4.3.1.

Nous construirons deux tableaux. Dans le premier chaque individu est décrit par la moyenne de l'essai auquel il appartient (Tableau 6.9). Dans le second, chaque individu est décrit par l'écart à la moyenne de l'essai auquel il appartient (Tableau 6.14). On effectue les ACP de ces fichiers sans réduire les variables (ACP sur matrice de variance).

Résultats sur le fichier des moyennes:

Tableau 6.7: Caractéristique des variables

Variable	Moyenne	Variance	Écart-type	Minimum	Maximum	Nombre d'individus	Essai
T_1	0.623	1.066	1.032	-1.262	2.507	20 ind.	1
T_2	0.744	0.859	0.927	-0.425	2.799	20 ind.	1
T_3	0.504	1.211	1.100	-1.466	2.216	20 ind.	1
T_4	0.880	0.988	0.994	-0.940	3.177	20 ind.	1
T_5	0.696	0.756	0.869	-0.755	3.103	20 ind.	1
T_6	0.342	1.240	1.114	-1.379	2.963	20 ind.	1
T_1	-0.296	0.543	0.737	-2.205	0.623	20 ind.	2
T_2	-0.404	0.674	0.821	-2.036	0.784	20 ind.	2
T_3	-0.141	0.697	0.835	-1.466	1.480	20 ind.	2
T_4	-0.236	0.258	0.508	-1.157	0.793	20 ind.	2
T_5	-0.338	0.741	0.861	-1.681	1.097	20 ind.	2
T_6	0.131	0.805	0.897	-1.642	1.773	20 ind.	2
T_1	-0.408	0.728	0.853	-2.676	0.623	16 ind.	3
T_2	-0.425	0.507	0.712	-1.231	0.784	16 ind.	3
T_3	-0.454	0.568	0.754	-1.466	1.480	16 ind.	3
T_4	-0.805	0.258	0.507	-2.241	-0.074	16 ind.	3
T_5	-0.446	0.682	0.826	-1.527	0.942	16 ind.	3
T_6	-0.591	0.427	0.653	-1.510	0.460	16 ind.	3

Tableau 6.8: Descriptions des individus supplémentaires

Individu	Axe 1	CTR	Axe 2	CTR	Axe 3	CTR
57	1.03	148	-0.92	119	-1.64	374
58	-0.97	148	0.86	119	1.53	374
59	0.92	202	-0.26	16	0.79	150
60	-1.26	202	0.36	16	-1.08	150
61	0.59	163	0.74	254	-0.46	97
62	-0.89	163	-1.12	254	0.69	97

Individu	Axe 1	CTR	Axe 2	CTR	Axe 3	CTR
63	0.90	160	-1.00	200	-0.40	32
64	-1.27	160	-1.42	200	0.57	32
65	0.73	190	-0.25	22	0.99	347
66	-1.35	190	0.46	22	-1.83	347
67	0.61	136	1.02	389	0.01	0
68	-0.99	136	-1.68	389	-0.01	0

Figure 6.3: Représentation des 12 individus supplémentaires sur les deux premiers axes

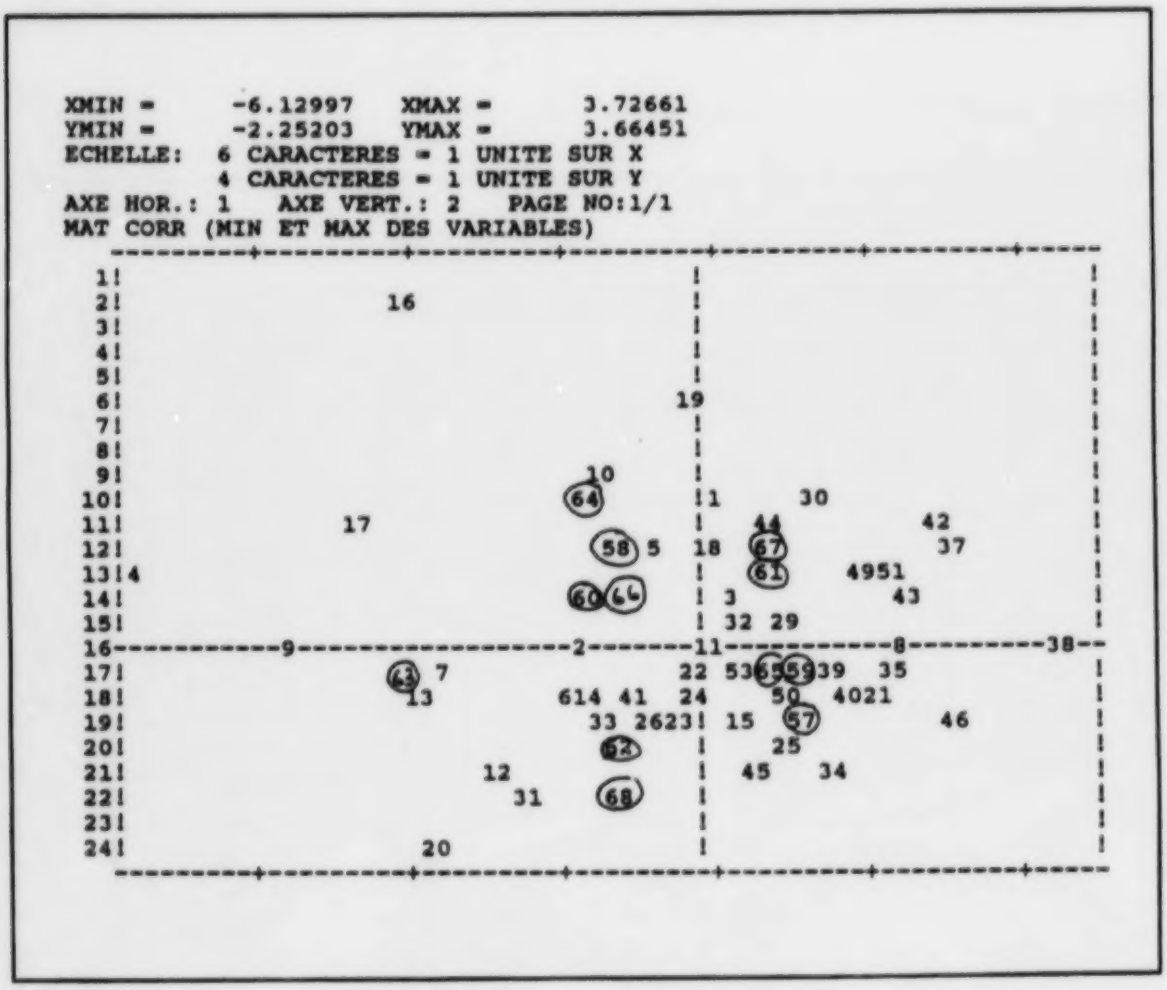


Tableau 6.9: Répétition des moyennes des essais

Individu	T_1	T_2	T_3	T_4	T_5	T_6
1	0.6230	0.7440	0.5040	0.8800	0.6960	0.3420
2	0.6230	0.7440	0.5040	0.8800	0.6960	0.3420
⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	0.6230	0.7440	0.5040	0.8800	0.6960	0.3420
20	0.6230	0.7440	0.5040	0.8800	0.6960	0.3420
21	-0.2960	-0.4040	-0.1410	-0.2360	-0.3380	0.1310
22	-0.2960	-0.4040	-0.1410	-0.2360	-0.3380	0.1310
⋮	⋮	⋮	⋮	⋮	⋮	⋮
39	-0.2960	-0.4040	-0.1410	-0.2360	-0.3380	0.1310
40	-0.2960	-0.4040	-0.1410	-0.2360	-0.3380	0.1310
41	-0.4080	-0.4250	-0.4540	-0.8050	-0.4460	-0.5910
42	-0.4080	-0.4250	-0.4540	-0.8050	-0.4460	-0.5910
⋮	⋮	⋮	⋮	⋮	⋮	⋮
55	-0.4080	-0.4250	-0.4540	-0.8050	-0.4460	-0.5910
56	-0.4080	-0.4250	-0.4540	-0.8050	-0.4460	-0.5910

Tableau 6.10: Matrice des covariances

	T_1	T_2	T_3	T_4	T_5	T_6
T_1	0.217					
T_2	0.258	0.308				
T_3	0.180	0.209	0.157			
T_4	0.315	0.366	0.275	0.482		
T_5	0.243	0.288	0.200	0.350	0.271	
T_6	0.131	0.144	0.132	0.232	0.145	0.148

Tableau 6.11: Description des variables actives

Individu	Axe 1	CTA	CTR	Axe 2	CTA	CTR
T_1	-0.46	142	982	0.06	49	18
T_2	-0.54	195	954	0.12	177	46
T_3	-0.39	103	992	-0.04	16	8
T_4	-0.69	317	990	-0.07	63	10
T_5	-0.51	176	978	0.08	74	22
T_6	-0.31	66	667	-0.22	620	333
Moyenne			927.2			72.8
Moyenne cumulée			927.2			1000.0

Tableau 6.12: Édition des premières valeurs propres

Somme des valeurs propres: 1.58172
Somme des carrés des val. pr.: 2.26349

Val. prop.	Inertie(%)	% cumulés
1.50240	94.99	94.99
0.07932	5.01	100.00
0.00000	0.00	100.00
0.00000	0.00	100.00
0.00000	0.00	100.00
0.00000	0.00	100.00

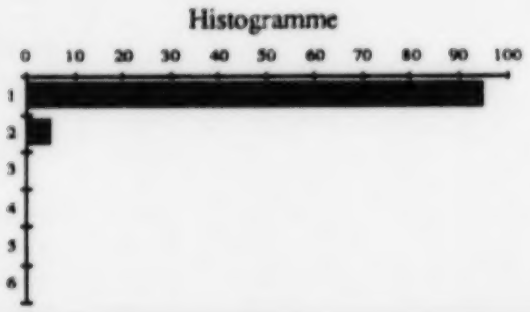


Tableau 6.13: Description des individus actifs.
Deux premières composantes principales normées aux valeurs propres correspondantes

Individu	CP1	CTA	CTR	CP2	CTA	CTR
1	-1.60	30	997	0.09	2	3
2	-1.60	30	997	0.09	2	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	-1.60	30	997	0.09	2	3
20	-1.60	30	997	0.09	2	3
21	0.58	4	727	-0.35	28	273
22	0.58	4	727	-0.35	28	273
⋮	⋮	⋮	⋮	⋮	⋮	⋮
39	0.58	4	727	-0.35	28	273
40	0.58	4	727	-0.35	28	273
41	1.28	19	936	0.33	25	64
42	1.28	19	936	0.33	25	64
⋮	⋮	⋮	⋮	⋮	⋮	⋮
55	1.28	19	936	0.33	25	64
56	1.28	19	936	0.33	25	64

Résultats sur le fichier des écarts aux moyennes:

Tableau 6.14: Données centrées-réduites recentrées par rapport aux moyennes des essais

T_1	T_2	T_3	T_4	T_5	T_6	T_1	T_2	T_3	T_4	T_5	T_6
-0.472	-1.169	-0.866	0.563	-0.834	-1.458	-0.495	0.382	-0.957	-0.271	0.818	-0.328
-0.472	-0.363	0.239	-0.303	0.092	0.118	0.447	1.188	-1.325	-0.704	0.509	-1.773
-0.472	-1.169	-0.866	-0.520	-0.988	-0.670	0.919	1.188	1.621	-0.054	0.663	1.642
1.413	2.055	1.712	2.080	2.407	1.431	0.447	0.382	-0.221	0.162	0.046	-0.328
-0.472	-0.363	-0.866	-0.087	-0.216	-0.670	0.447	1.188	1.252	0.162	0.509	0.591
-0.472	-0.363	0.607	-0.087	-0.371	0.643	-1.437	-0.021	-0.221	-0.921	0.046	0.460
0.000	1.249	0.976	-0.087	0.092	0.774	-0.495	-1.230	-0.221	0.379	-1.343	-0.459
-1.885	-1.169	-1.602	-1.820	-1.451	-1.064	0.447	0.785	0.516	0.596	1.280	1.117
1.413	1.249	1.344	0.563	1.018	1.431	-0.024	-0.827	-1.325	-0.271	-1.189	-1.510
0.942	0.040	-0.866	0.130	-0.216	-1.196	-1.909	-1.230	-1.325	-0.704	-1.189	-1.510
-0.472	-0.766	-0.129	-0.737	-1.142	-0.539	-0.495	-0.827	-0.957	-0.271	0.509	0.197
-0.472	0.443	0.976	-0.954	0.401	1.168	0.447	-0.424	0.516	-0.704	1.343	-0.722
0.942	0.846	1.344	-0.303	0.092	0.774	0.559	0.806	1.197	0.298	1.080	0.919
-0.943	-0.363	0.239	-0.303	0.092	0.643	0.559	-0.806	-1.012	0.298	-1.081	-0.919
-1.885	-1.169	-0.497	-0.954	-0.371	0.249	0.088	-0.403	-0.644	-0.352	-0.463	-0.394
1.884	0.846	-1.234	2.297	1.172	-1.196	-0.383	1.209	-0.644	0.515	1.388	-0.919
1.413	0.846	0.607	1.213	1.018	0.118	1.031	-0.806	1.934	0.081	-0.772	0.788
-0.472	-0.766	-0.866	-0.520	-0.216	-1.196	-2.268	0.806	-0.276	-1.436	0.154	-0.394
0.000	-0.363	-1.970	0.780	-0.216	-1.721	-0.383	-0.806	0.092	0.081	-0.618	0.000
0.471	0.443	1.712	-0.954	-0.371	2.351	0.559	-0.000	-0.276	-0.352	-1.081	-0.657
-0.024	-1.632	0.148	0.162	-1.343	-0.197	1.031	-0.000	-0.644	-0.352	-0.309	-0.394
0.447	-0.021	0.148	0.596	0.046	0.591	0.088	0.806	0.461	-0.352	-0.155	0.394
0.447	-0.021	0.884	0.162	0.200	0.460	0.559	-0.403	-0.644	-0.135	-0.463	-0.657
-0.024	0.382	0.148	-0.054	0.818	0.591	0.559	-0.806	-0.276	0.731	-0.618	-0.263
-0.495	-0.424	0.516	-0.271	-0.417	0.329	0.559	-0.000	0.829	0.081	0.308	0.394
0.919	-0.021	0.884	0.596	-0.263	0.985	-0.854	-0.403	0.092	-0.135	0.000	0.131
0.919	0.785	0.148	1.029	0.200	-0.722	-0.854	-0.403	0.092	0.515	1.234	1.051
-0.495	0.382	-0.221	0.379	1.435	0.591	-0.854	1.209	-0.276	0.515	1.388	0.919

Tableau 6.15: Matrice des covariances

	T_1	T_2	T_3	T_4	T_5	T_6
T_1	0.783					
T_2	0.322	0.692				
T_3	0.318	0.337	0.843			
T_4	0.404	0.208	0.096	0.518		
T_5	0.194	0.560	0.237	0.312	0.729	
T_6	0.140	0.323	0.720	0.056	0.365	0.852

Tableau 6.16: Édition des premières valeurs propres

Somme des valeurs propres: 4.41868
Somme des carrés des val. pr.: 6.9645

Val. prop.	Inertie(%)	% cumulés
2.33374	52.82	52.82
0.98873	22.38	75.19
0.66933	15.15	90.34
0.28523	6.46	96.79
0.09524	2.16	98.95
0.04641	1.05	100.00

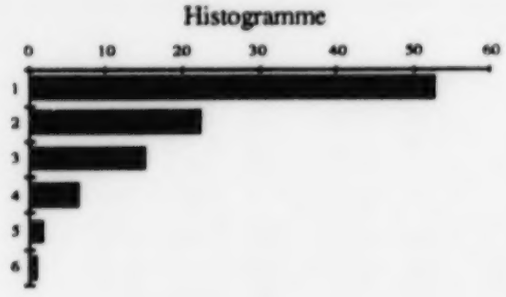


Tableau 6.17: Description des variables actives

	Axe 1	CTA	CTR	Axe 2	CTA	CTR	Axe 3	CTA	CTR
T_1	-0.53	120	358	-0.45	209	264	0.51	383	328
T_2	-0.66	188	634	-0.20	42	59	-0.30	135	130
T_3	-0.74	234	646	0.43	188	220	0.27	112	89
T_4	-0.37	57	258	-0.48	234	446	0.12	20	26
T_5	-0.64	176	562	-0.23	52	71	-0.48	349	321
T_6	-0.72	225	616	0.52	276	320	-0.01	0	0
Moyenne			512.5			230.0			149.0
Moyenne cumulée			512.5			742.5			891.5

Tableau 6.18: Description des individus actifs

	CP1	CTA	CTR	CP2	CTA	CTR	CP3	CTA	CTR
1	1.99	30	727	-0.77	11	108	0.41	9	34
2	0.18	0	44	0.98	6	642	-0.18	1	99
3	1.94	29	938	0.20	1	10	0.36	4	33
4	-4.40	148	906	-1.13	23	60	-0.44	5	9
5	1.17	10	850	-0.35	2	74	-0.33	3	67
6	-0.10	0	8	1.02	19	807	0.25	2	48
7	-1.40	15	636	0.60	4	113	-0.20	1	13
8	3.48	93	864	1.06	20	98	-0.67	12	32
9	-2.92	65	974	-0.07	0	1	0.33	3	13
10	0.79	4	157	-1.65	38	674	0.43	5	60
11	1.47	16	728	0.65	8	143	0.52	7	92
12	-1.00	8	260	1.53	42	616	-0.51	7	68
13	-1.68	22	684	0.81	5	84	0.62	10	94
14	0.10	0	6	1.07	21	722	-0.47	6	140
15	1.47	21	442	1.56	44	389	-0.82	18	108
16	-0.90	4	58	-3.58	231	921	0.08	0	0
17	-1.93	28	880	-1.31	31	308	0.34	3	20
18	1.70	22	871	-0.33	2	37	-0.24	2	17
19	1.83	26	439	-3.01	73	530	-0.28	2	10
20	-1.92	28	370	2.22	89	495	0.77	16	80
21	1.26	12	350	0.83	5	63	1.49	94	663
22	-0.66	3	469	-0.13	0	17	0.39	4	161
23	-0.91	6	664	0.30	1	71	0.48	6	184
24	-0.84	5	592	0.15	0	18	-0.60	10	304
25	0.19	0	34	0.93	16	936	0.23	1	50
26	-1.34	13	807	0.26	1	33	1.10	33	404
27	-0.73	4	166	-1.44	37	667	0.36	3	42
28	-0.88	6	247	-0.15	0	7	-1.32	46	582
29	0.35	1	95	-0.49	4	113	-1.29	44	746
30	0.77	4	81	-1.73	54	411	-0.99	26	135
31	-2.66	54	883	0.79	11	79	0.26	3	9
32	-0.12	0	26	-0.64	7	770	0.06	0	7
33	-1.61	25	857	0.21	1	12	-0.03	0	0
34	0.60	3	112	1.24	38	485	-1.13	33	393
35	1.50	17	589	0.26	1	17	0.92	33	214
36	-1.95	29	883	-0.14	0	4	-0.53	7	63
37	3.29	40	842	-0.79	11	99	0.52	7	43
38	3.23	79	933	0.37	2	12	-0.56	8	28
39	0.75	4	284	0.10	0	4	-0.66	13	199
40	0.85	6	210	0.37	3	46	1.30	48	491
41	-3.08	33	984	0.19	1	8	-0.15	1	5
42	1.46	16	523	-0.91	15	203	0.99	26	340
43	0.92	6	786	-0.17	1	26	0.21	1	42
44	-0.38	1	34	-1.40	35	387	-1.64	73	931
45	-1.01	8	193	1.08	21	175	2.04	111	624
46	1.04	8	132	1.21	26	179	-2.08	116	936
47	0.68	4	388	0.48	4	194	0.47	6	182
48	0.79	5	294	-0.30	2	43	0.85	19	340
49	0.35	1	88	-0.73	9	277	0.56	8	168
50	-0.64	3	350	0.41	3	141	-0.05	0	2
51	0.83	5	443	-0.63	7	252	0.54	8	184
52	0.50	2	123	-0.96	6	156	1.02	28	514
53	-0.93	7	689	0.30	1	32	0.45	5	161
54	0.40	1	168	0.68	8	447	-0.37	4	146
55	-0.71	4	136	0.53	5	75	-1.01	27	271
56	-1.24	12	288	-0.04	0	1	-1.82	88	824
MOYENNE			444.5			334.5			180.2
MOYENNE CUMULEE			444.5			479.0			859.2

On remarquera d'abord que l'addition des deux matrices redonne la matrice de corrélation initiale ce qui entraîne que la somme des deux sommes des valeurs propres est égale à 6=1.58172 + 4.41868. On a décomposé l'inertie totale en deux parts additives, l'une due aux essais, l'autre étant une inertie résiduelle au sein des essais. On doit remarquer que l'inertie due aux essais n'est pas négligeable.

L'inertie du fichier des moyennes n'a que deux valeurs propres non nulles ce qui n'est pas étonnant puisqu'il n'y a que trois essais. Seule la première valeur propre est importante. Elle est associée à un facteur principal très semblable au premier facteur de l'étude initiale. On retrouve ce facteur dans l'ACP des écarts à la moyenne ce qui prouve que les essais à eux seuls ne déterminent pas l'intensité des scores.

6.3 Élimination de l'effet taille.

On associe à chaque individu la moyenne des six variables qui lui sont associées. Un calcul

classique montre que la variance de ces moyennes par individu est égale à 3,685. On construit par ailleurs le fichier des écarts de chaque ensemble à la moyenne de l'individu correspondant. L'ACP de ce fichier (sur la matrice des variances) donnent les résultats suivants:

Tableau 6.19: Données centrées-par rapport à la moyenne de l'individu

T_1	T_2	T_3	T_4	T_5	T_6	T_1	T_2	T_3	T_4	T_5	T_6
0.226	-0.350	-0.287	1.517	-0.063	-1.042	-0.435	0.334	-0.742	-0.151	0.836	0.159
-0.366	-0.136	0.226	0.060	0.271	-0.057	0.641	1.274	-0.976	-0.450	0.661	-1.152
0.300	-0.276	-0.213	0.509	-0.143	-0.179	-0.160	0.001	0.697	-1.072	-0.458	0.990
-0.445	0.318	-0.265	0.479	0.622	-0.708	0.284	0.111	-0.229	0.059	-0.159	-0.064
-0.035	0.195	-0.548	0.607	0.294	-0.514	-0.327	0.306	0.633	-0.552	-0.307	0.244
-0.473	-0.243	0.487	0.169	-0.299	0.361	-1.170	0.138	0.201	-0.594	0.271	1.154
-0.509	0.861	0.348	-0.339	-0.344	-0.016	-0.016	-0.858	0.414	0.919	-0.905	0.447
-0.395	0.442	-0.231	-0.073	0.112	0.145	-0.425	-0.195	-0.201	-0.216	0.366	0.672
0.235	0.192	0.047	-0.358	-0.087	-0.028	0.752	-0.159	-0.394	0.565	-0.455	-0.307
1.128	0.347	-0.799	0.573	0.043	-1.291	-0.680	-0.109	0.059	0.585	-0.002	0.146
0.150	-0.023	0.374	0.142	-0.447	-0.198	-0.270	-0.710	-0.577	0.014	0.692	0.849
-0.741	0.295	0.588	-0.966	0.205	0.618	0.737	-0.242	0.961	-0.354	-1.095	-0.005
0.318	0.343	0.601	-0.670	-0.459	-0.131	-0.137	0.093	0.455	-0.795	0.346	0.040
-0.846	-0.145	0.217	0.051	0.262	0.459	1.166	-0.216	-0.451	0.508	-0.512	-0.495
-1.122	-0.285	0.147	0.066	0.465	0.731	0.563	0.055	-0.215	-0.274	-0.026	-0.102
1.247	0.330	-1.990	1.917	0.608	-2.114	-0.464	1.111	-0.771	0.037	1.269	-1.183
0.535	0.089	-0.390	0.592	0.213	-1.041	0.769	-1.086	1.625	-0.578	-1.073	0.342
0.192	0.019	-0.321	0.401	0.521	-0.813	-1.586	1.472	0.360	-1.150	0.799	0.105
0.573	0.331	-1.516	1.610	0.430	-1.429	0.003	-0.437	0.432	0.070	-0.270	0.203
-0.146	-0.053	0.976	-1.314	-0.915	1.453	0.974	0.398	0.093	-0.334	-0.704	-0.425
0.375	-1.341	0.702	0.621	-0.986	0.629	1.256	0.208	-0.465	-0.524	-0.122	-0.352
0.064	-0.512	-0.080	0.273	-0.379	0.635	-0.005	0.696	0.322	-0.842	-0.287	0.118
0.010	-0.566	0.602	-0.215	-0.279	0.450	0.963	-0.016	-0.286	-0.128	-0.097	-0.436
-0.416	-0.118	-0.089	-0.386	0.384	0.626	0.785	-0.597	-0.096	0.560	-0.430	-0.220
-0.450	-0.487	0.716	-0.166	-0.414	0.801	0.311	-0.265	0.535	-0.564	0.022	-0.037
0.320	-0.728	0.440	0.057	-0.904	0.813	-0.546	-0.112	0.354	-0.224	0.270	0.256
0.444	0.202	-0.172	0.614	-0.317	-0.770	-1.013	-0.579	-0.113	-0.041	1.037	0.709
-0.922	-0.153	-0.493	0.012	0.966	0.591	-1.224	0.822	-0.692	-0.252	0.980	0.366

Tableau 6.20: Matrice des covariances

	T_1	T_2	T_3	T_4	T_5	T_6
T_1	0.446					
T_2	-0.059	0.276				
T_3	-0.089	-0.126	0.380			
T_4	0.147	-0.083	-0.234	0.410		
T_5	-0.182	0.144	-0.214	0.025	0.317	
T_6	-0.263	-0.153	0.284	-0.264	-0.090	0.486

Tableau 6.21: Édition des premières valeurs propres

Somme des valeurs propres: 2.31479
Somme des carrés des val. pr.: 1.8483

Val. prop.	Inertie(%)	% cumulés
1.11792	48.29	48.29
0.68238	29.48	77.77
0.34120	14.74	92.51
0.11378	4.92	97.43
0.05951	2.57	100.00
0.00000	0.00	100.00

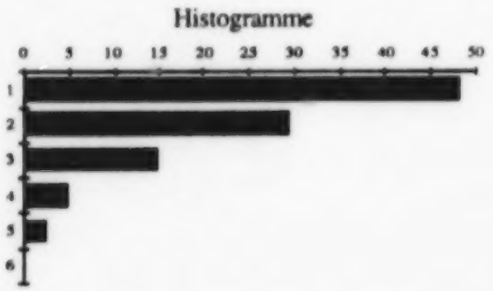


Tableau 6.22: Description des variables actives

	Axe 1	CTA	CTR	Axe 2	CTA	CTR	Axe 3	CTA	CTR
T1	-0.37	122	306	0.49	352	538	-0.21	129	98
T2	-0.15	21	84	-0.34	173	427	-0.34	331	409
T3	0.52	245	721	0.23	78	140	-0.09	21	19
T4	-0.48	207	564	0.14	29	48	0.38	412	343
T5	-0.17	26	93	-0.50	368	793	0.09	21	23
T6	0.65	379	872	-0.02	0	1	0.17	85	60
Moyenne			440.0			324.4			158.7
Moyenne cumulée			440.0			764.4			923.2

On voit que la somme des valeurs propres additionnée à la variance des moyennes redonne 6. Il est donc clair que l'effet taille est très important dans cette étude. Le facteur principal 1 est très semblable au facteur principal 2 de l'ACP initiale: c'est bien logique puisque la composante 1 dans l'ACP initiale était une variable de taille. La considération du graphique des individus permet alors de reconnaître des individus qui, indépendamment de leur score total, ont eu des comportements d'amélioration ou de détérioration semblables.

Tableau 6.23: Description des individus actifs

	CP1	CTA	CTR	CP2	CTA	CTR	CP3	CTA	CTR
1	-1.49	34	610	0.51	7	73	0.82	35	186
2	0.15	0	83	-0.35	1	222	0.24	3	198
3	-0.49	4	458	0.41	4	321	0.34	6	215
4	-0.78	10	405	-0.79	15	381	0.21	2	29
5	-0.93	14	612	-0.32	3	97	0.26	4	65
6	0.44	6	525	0.16	1	33	0.41	9	217
7	0.43	3	134	-0.41	4	125	-0.64	21	299
8	0.06	0	9	-0.97	9	721	-0.07	0	10
9	0.07	0	33	0.07	0	19	-0.48	11	887
10	-1.90	58	898	0.40	4	39	-0.49	13	60
11	0.02	0	1	0.86	7	999	-0.13	1	38
12	1.29	27	716	-0.70	13	208	-0.40	8	48
13	0.44	3	181	0.38	4	116	-0.94	46	696
14	0.64	7	388	-0.84	8	274	0.96	16	294
15	0.85	12	342	-0.79	16	294	0.87	39	356
16	-3.74	223	989	0.04	0	0	0.36	7	9
17	-1.34	29	939	0.16	1	14	-0.08	0	3
18	-1.00	16	804	-0.23	1	38	0.06	0	3
19	-2.88	116	981	-0.10	1	4	0.90	13	34
20	2.18	76	942	0.51	7	47	-0.61	20	66
21	0.67	7	108	1.67	73	664	0.97	80	224
22	0.34	2	129	0.49	4	271	0.99	18	387
23	0.80	10	487	0.83	7	382	0.19	2	38
24	0.82	6	431	-0.83	7	333	0.33	3	99
25	1.22	24	827	0.34	3	66	0.40	9	91
26	0.83	11	500	1.16	15	981	0.38	9	63
27	-0.97	15	706	0.44	5	146	-0.13	1	13
28	0.30	1	38	-1.32	19	617	0.61	38	275
29	-0.23	1	34	-1.14	14	813	0.14	1	13
30	-1.80	34	455	-0.88	20	599	-1.35	98	360
31	1.87	39	867	0.18	1	11	-0.91	14	92
32	-0.27	1	406	0.17	1	155	-0.14	1	104
33	0.83	11	457	-0.06	0	3	-0.48	12	217
34	1.43	33	630	-0.98	23	303	0.31	3	29
35	0.34	2	41	1.14	35	485	1.03	88	381
36	0.83	4	315	-0.90	7	280	0.48	9	184
37	-0.61	10	477	0.78	16	444	0.09	0	9
38	0.11	0	13	-0.24	2	71	0.72	37	609
39	0.32	2	47	-0.46	4	180	0.98	47	426
40	0.99	6	121	1.41	53	699	-0.65	22	150
41	0.89	4	383	-0.34	3	116	-0.82	14	273
42	-1.05	18	467	1.06	29	474	-0.12	1	6
43	-0.25	1	133	0.23	1	109	-0.41	9	273
44	-1.33	28	390	-1.69	75	567	-0.80	13	69
45	1.34	29	398	1.91	95	685	-0.32	5	17
46	0.98	15	143	-2.13	110	671	-0.92	45	136
47	0.41	3	343	0.48	4	484	0.28	3	138
48	-0.35	2	64	0.82	17	381	-1.03	56	961
49	-0.66	7	192	0.82	7	120	-0.96	48	409
50	0.96	8	237	-0.17	1	22	-0.99	52	709
51	-0.67	7	367	0.84	8	241	-0.52	14	219
52	-0.56	5	203	1.05	29	717	0.31	5	62
53	0.42	3	232	0.34	3	146	-0.41	9	218
54	0.60	6	373	-0.39	4	237	0.18	2	81
55	0.67	7	152	-1.04	28	388	1.04	57	389
56	0.18	0	4	-1.91	95	953	0.16	1	4
MOYENNE			381.3			290.1			197.0
MOYENNE CUMULEE			381.3			672.4			869.4

Références

[1] BONIFAS L., ESCOUFIER Y., GONZALEZ P.L., SABATIER R. (1984): Choix des variables en Analyse en Composantes Principales. R.S.A., Vol. XXXII, nº 2, p. 5-15.

[2] CHESSEL D., LEBRETON J.D., YOCCOZ N. (1987): Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. Rev. Stat. Appli., 35, p. 55-72.

[3] GOWER J.C., HARDING S.A. (1988): Non-linear Biplots. À paraître.

[4] HOTELLING H. (1933): Analysis of a complex of statistical variables in principal components. J. Educ. Psychol., 24, p. 417-441 et 498-520.

[5] LAVIT C. (1988): L'analyse conjointe de plusieurs tableaux de données. Masson.

[6] LEBART L. (1969): Analyse statistique de la contiguïté. Pub. Inst. Stat. Univ., Paris, p. 81-112.

[7] LEBART L., FÉNELON J.-P. (1971): Statistique et Informatique Appliquées. Dunod. Chap. IV.4.

[8] LEBRETON J.D., CHESSEL D., PRODON R., YOCCOZ N. (1988): L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. 1–Variable de milieu quantitative. 2–Variable de milieu qualitative. Acta Oecologica, Oecologica Generalis, 9, sous presse.

[9] MOM A. (1988): Méthodologie Statistique de Classification dans les Réseaux de Transport. Thèse de Doctorat. Université des Sciences et Techniques du Languedoc. Montpellier.

[10] OKAMOTO M. (1969): Optimality of principal component. Mult. An. II, North Holland Publishing Company, p. 673-685.

[11] PEARSON K. (1901): On lines and planes of closest fit to system of points in spaces. Phil. Mag., 2, 6, p. 559-572.

[12] RAO C. R. (1964): The use and interpretation of principal component analysis in applied research. Sankyā, A, 26, p. 329-359.

[13] RAO C. R. (1979): Separation theorems for singular values of matrices and their application in Multivariate Analysis. J. of Mult. An., 9, p. 362-377.

[14] SABATIER R. (1987): Méthodes factorielles en Analyse des données: Approximations et prises en compte de variables concomitantes. Thèse de Doctorat d'État. Université des Sciences et Techniques du Languedoc. France.

[15] SCHEFFÉ H. (1959): The Analysis of Variance. John-Wiley and Sons.

Annexe 1: Scores Échelle d'Hamilton

T_1 - Total partiel à la période 1
T_2 - Total partiel à la période 2
T_3 - Total partiel à la période 3
T_4 - Total à la période 1
T_5 - Total à la période 2
T_6 - Total à la période 3

T_1	T_2	T_3	T_4	T_5	T_6
7	4	3	28	15	4
7	6	6	24	21	16
7	4	3	23	14	10
11	12	10	35	36	26
7	6	3	25	19	10
7	6	7	25	18	20
8	10	8	25	21	21
4	4	1	17	11	7
11	10	9	28	27	26
10	7	3	26	19	6
7	5	5	22	13	11
7	8	8	21	23	24
10	9	9	24	21	21
6	6	6	24	21	20
4	4	4	21	18	17
12	9	2	36	28	6
11	9	7	31	27	16
7	5	3	23	19	6
8	6	0	29	19	2
9	8	10	21	18	33
6	0	4	21	5	12
7	4	4	23	14	18
7	4	6	21	15	17
6	5	4	20	19	18
5	3	5	19	11	16
8	4	6	23	12	21
8	6	4	25	15	8
5	5	3	22	23	18

T_1	T_2	T_3	T_4	T_5	T_6
5	5	1	19	19	11
7	7	0	17	17	0
8	7	8	20	18	26
7	5	3	21	14	11
7	7	7	21	17	18
3	4	3	16	14	17
5	1	3	22	5	10
7	6	5	23	22	22
6	2	0	19	6	2
2	1	0	17	6	2
5	2	1	19	17	15
7	3	5	17	5	8
7	6	6	19	20	15
7	2	0	19	6	1
6	3	1	16	10	5
5	7	1	20	22	1
8	2	8	18	8	14
1	6	2	11	14	5
5	2	3	18	9	8
7	4	2	16	6	3
8	4	1	16	11	5
6	6	4	16	12	11
7	3	1	17	10	3
7	2	2	21	9	6
7	4	5	18	15	11
4	3	3	17	13	9
4	3	3	20	21	16
4	7	2	20	22	15

Annexe 2: Structuration des individus

X^1 - Lot thérapeutique X^2 - Essai X^3 - Sexe

X^1	X^2	X^3
2	1	1
2	1	1
1	1	1
2	1	2
2	1	2
1	1	1
1	1	1
1	1	1
1	1	1
2	1	2
2	1	2
1	1	2
1	1	2
1	1	2
1	1	2
2	1	2
2	1	1
1	1	2
1	1	1
1	1	2
1	2	2
2	2	2
1	2	1
1	2	2
1	2	2
1	2	2
1	2	2
2	2	1

X^1	X^2	X^3
1	2	1
2	2	2
1	2	2
1	2	2
1	2	1
1	2	1
1	2	1
2	2	1
1	2	1
1	2	1
2	2	2
1	2	2
2	3	1
2	3	1
1	3	2
1	3	2
2	3	1
1	3	2
1	3	2
1	3	2
1	3	1
2	3	1
2	3	1
2	3	1
2	3	1
2	3	1
1	3	1
1	3	1

La Classification Automatique

Edwin Diday
Paris IX, Dauphine

1 Les premières étapes d'une analyse de données

1.1 Les principaux choix de base

1.1.1 Choix de l'ensemble des individus

Il existe d'innombrables façons de définir un ensemble d'individus. Les techniques de recueil de données diffèrent, bien sûr, d'un domaine à l'autre, voire dans le même domaine. Elles diffèrent suivant le type de problème posé. Il se produit parfois que l'ensemble des individus traités soit exhaustif mais dans le cas le plus fréquent, il s'agit d'un échantillon d'une population plus vaste. La façon dont les individus ont été recueillis est alors fondamentale et on peut faire appel à la théorie des sondages et aux pratiques qui en résultent : plan de sondage, sondage aléatoire (élémentaire, en groupe, etc ...) ou plus empiriquement à la méthode des quotas (pour plus de détails voir par exemple GROSBRAS (1987)).

Dans la pratique, l'analyste de données intervient généralement seulement en aval du recueil et utilise des méthodes de description qui sont souvent indépendantes de la façon dont les données ont été obtenues: en effet, beaucoup de méthodes d'analyse des données ne nécessitent pas d'hypothèse sur la façon dont les individus ont été recueillis et ont pour objectif essentiel de traiter les données telles qu'elles sont fournies. Dans toute la suite, on note Ω l'ensemble des individus.

1.1.2 Choix de l'ensemble des variables

À chaque paramètre choisi par l'utilisateur pour décrire les individus, on peut associer une ou plusieurs variables.

Le choix des variables est étroitement lié au problème posé. C'est essentiellement au spécialiste que revient cette tâche. Cependant l'analyse des données peut permettre d'en vérifier la validité en faisant ressortir, par exemple, les redondances et la pertinence de certaines variables.

Après avoir choisi les variables, il faut leur associer un "type". On peut distinguer deux grands types de variables: les variables **quantitatives** et les variables **qualitatives**.

Les variables quantitatives: L'analyse des données permettra entre autres de mettre en évidence des liaisons linéaires entre de telles variables. Dans la pratique, on distingue les types suivants:

- quantitatif mesurable (revenu, poids, ...)
- quantitatif d'ordre (note, rang, ...)
- quantitatif de comptage (fréquence, contingence, ...)
- quantitatif binaire (succès-échec, présence-absence, ...)

Les variables qualitatives: Les classes d'une variable qualitative sont appelées modalités de la variable. On distingue essentiellement les types suivants:

- qualitatif nominal (lieu géographique, catégorie socio-professionnelle, ...).
- qualitatif ordinal (pas d'accord, sans opi-

Le présent texte est un extrait d'une version préliminaire des deux premiers chapitres du livre de C. CELEUX et al., *La classification automatique des données.*

Ecole d'été sur l'analyse des données, juin 1988, UQAM, SCAD-CIRADE.

nion, tout à fait d'accord, ...).
- qualitatif textuel (titre de film, nom d'auteur, ...).

Une variable v étant choisie, le codage intervient chaque fois que l'on désire modifier les valeurs prises par v. Ainsi, réaliser un codage revient à effectuer un changement de variables.

Exemple: Considérons v la variable "âge": c'est une variable quantitative mesurable. Si l'on désire transformer l'âge en classes d'âge, on obtient une nouvelle variable v' qui est qualitative ordinale. Si maintenant, on ne désire plus faire intervenir l'ordre entre les tranches d'âge, on obtient une variable qualitative nominale.

Nous traiterons en détail la notion de codage au paragraphe 1.4. Il s'agit d'un choix de base fondamental en analyse des données, car on est souvent amené à transformer les variables pour les rendre plus homogènes entre elles ou pour les rendre compatibles avec des méthodes qui leur donnent un éclairage enrichissant. Par exemple, on transforme souvent des variables quantitatives en variables qualitatives afin de faire ressortir des liaisons non linéaires. L'analyste de données dispose d'une grande gamme d'outils permettant de passer d'un type de variable à un autre. Des choix qu'il fera, dépendra le résultat de l'analyse.

1.1.3 Choix d'une similarité ou d'une dissimilarité

Il s'agit d'un choix important qui intervient dans la plupart des méthodes de classification automatique ou d'analyse factorielle. Cette notion vise essentiellement à associer une valeur numérique à chaque couple d'individus ou de variables pour mesurer leur ressemblance. Nous présenterons au paragraphe 1.5 les principaux choix possibles: les indices de similarité, de dissimilarité, de distance, les ultramétriques etc...

1.2 Construction d'un tableau de données

Un tableau de données est un tableau rectangulaire qui se déduit de la définition de l'ensemble des individus et des variables. Soit n le nombre d'individus et p le nombre de variables. Notant $\{w_1, ..., w_n\}$ l'ensemble Ω des individus et $\{v_1, ..., v_p\}$ l'ensemble V des variables, le tableau de données associé est:

$$X = \left(x_i^j\right)_{i=1,\ ....\ n}^{j=1,\ ....\ p}$$

Tableaux individus × variables

Tableaux de données quantitatives:

Lorsque toutes les variables choisies sont quantitatives, on peut établir un tableau de données quantitatives qui contient des valeurs numériques quelconques. Les variables expriment, par exemple, des taux de production, des poids, des longueurs, etc....

Tableaux de données qualitatives ou de modalités:

Dans ce cas, toutes les variables sont qualitatives. Si elles sont toutes ordinales (respectivement nominales), on dira que l'on a un tableau de modalités ordonnées (respectivement non-ordonnées).

Tableaux binaires:

On rencontre souvent des tableaux qui ne contiennent que des 0 ou des 1. Ils sont appelés tableaux binaires. On les trouve par exemple dans des données d'enquêtes.

Tableaux de préférence:

Par exemple, on dispose des préférences de personnes interrogées sur des marques de parfum.

	Marques de parfum				
Personnes enquêtées	M_1	M_2	M_3	M_4	M_5
W_1	1	3	4	2	5
W_2	3	2	5	4	1
W_3	5	3	4	2	1
W_4	1	5	3	4	2

Ici Ω représente l'ensemble des personnes interrogées. Aux marques M_i sont associées des variables v_i: $\Omega \longrightarrow \{1, 2, 3, 4, 5\}$ qui peuvent être considérées comme un ensemble de variables qualitatives ordinales. $x_2^3 = 5$ signifie que le deuxième individu préfère la troisième marque de parfum à toutes les autres.

Tableaux hétérogènes:

Il s'agit de tableaux où les variables sont de différents types. Par exemple, au paramètre "prix", on associe la variable quantitative mesurable

$$v_1: \Omega \longrightarrow \mathbb{R}.$$

Au paramètre "mode de transport", on associe la variable qualitative nominale

$v_2: \Omega \longrightarrow$ {avion, bateau, train}.Au paramètre "fragilité", on associe la variable qualitative ordinale

$$v_3: \Omega \longrightarrow \{1,2,3\}.$$

Tableaux variables × variables

Tableaux de contingence et de fréquence:

À partir de deux variables qualitatives, il est habituel de définir un tableau croisant les modalités des deux variables de la manière suivante:

Connaissant les valeurs prises par les deux variables sur une population, il est simple d'associer à une modalité de la première variable et à une modalité de la seconde variable le cardinal de la sous-population ayant pris simultanément ces deux modalités. On obtient ainsi un tableau de nombres positifs dont le nombre de lignes est le nombre de modalités de la première variable et dont le nombre de colonnes est le nombre de modalités de la seconde variable. Ce tableau est appelé **tableau de contingence**. Si l'on divise chaque valeur de ce tableau par le cardinal de la population, on obtient un tableau de fréquences relatives que l'on appellera plus simplement **tableau de fréquences**.

Dans la suite de ce texte, on utilise les notations suivantes: I et J sont les deux variables qualitatives ayant respectivement n et p modalités. Afin de rendre plus lisible le texte, I et J représentent aussi l'ensemble des n modalités et l'ensemble des p modalités

$$I = \{1, \ldots, i, \ldots, n\}$$

$$J = \{1, \ldots, j, \ldots, p\}$$

où i est la $i^{ème}$ modalité de I et j la $j^{ème}$ modalité de J. n_{ij} est le nombre d'individus possédant à la fois la modalité i et la modalité j. Le tableau de contingence est l'ensemble $\{n_{ij}: i \in I, j \in J\}$. On pose

$$s = \sum_{i \in I} \sum_{j \in J} n_{ij}$$

s représente le cardinal de la population sur laquelle sont définies les deux variables qualitatives.

Le tableau F des fréquences est l'ensemble

$$\{f_{ij} = \frac{n_{ij}}{s}, i \in I, j \in J\}.$$

f_{ij} est une estimation de la probabilité qu'un individu présente simultanément la modalité i et la modalité j. On définit alors les fréquences marginales $f_{i.}$ et $f_{.j}$:

$$\forall i \in I,\ f_{i.} = \sum_{j \in J} f_{ij} \quad \text{et} \quad \forall j \in J,\ f_{.j} = \sum_{i \in I} f_{ij}.$$

Les vecteurs $(f_{1.}, \ldots, f_{i.}, \ldots, f_{n.})$ et $(f_{.1}, \ldots, f_{.j}, \ldots, f_{.p})$ sont notés f_I et f_J. Ce sont les lois marginales définies sur I et J.

On définit les fréquences conditionnelles f_j^i et f_i^j:

$$\forall i \in I,\ \forall j \in J,\ f_j^i = \frac{f_{ij}}{f_{i.}} \quad \text{et} \quad f_i^j = \frac{f_{ij}}{f_{.j}}$$

Les vecteurs $(f_{1.}^i, \ldots, f_{j.}^i, \ldots, f_{p.}^i)$ et $(f_{1.}^j, \ldots, f_{i.}^j, \ldots, f_{n.}^j)$ sont notés f_J^i et f_I^j. Ce sont les lois conditionnelles. Elles sont aussi appelées profils.

Tableau de contingence multiple:

Nous venons de voir qu'un tableau de contingence est construit à partir de deux variables qualitatives définies sur une même population. Ceci peut être généralisé en croisant non plus deux variables qualitatives mais deux familles de variables qualitatives définies sur une même population. On parle alors de tableau de contingence **multiple**.

Si on note $(I_1, I_2, ..., I_r)$ et $(J_1, J_2, ..., J_t)$ les deux familles de variables qualitatives, le tableau de contingence multiple associé est formé des tableaux de contingence de tous les couples (I_k, J_ℓ) avec $1 \leq k \leq r$ et $1 \leq \ell \leq t$.

Les deux familles ne sont pas nécessairement disjointes. Si tel est le cas les blocs obtenus en croisant une variable avec elle-même prennent alors la forme particulière d'une matrice diagonale dont chacun des éléments de la diagonale est le nombre d'occurrences de la modalité correspondante.

En particulier, si les deux familles de variables sont identiques, le tableau obtenu est appelé **tableau de Burt**. Il prend alors la forme d'un tableau carré symétrique dont les blocs diagonaux sont les matrices diagonales dont nous venons de parler.

Lorsqu'une des familles est réduite à une seule variable, le tableau obtenu est appelé **tableau de contingence juxtaposé**.

Tableaux de similitude:

Ces tableaux recensent des similitudes entre des variables. Par exemple, on considère souvent les tableaux de corrélation et de covariance définis par les corrélations ou les covariances de tous les couples de variables quantitatives.

Tableaux individus × individus:

Lorsqu'on évalue la similarité ou la dissimilarité entre chaque couple d'individus, on construit un **tableau de proximités**.

Tableaux à trois dimensions:

Enfin, l'utilisateur peut se trouver confronté à des tableaux à trois **dimensions**. Par exemple, si différents experts donnent leur avis, on obtient un ensemble de tableaux de proximités. On voit aussi une troisième dimension apparaître quand un même tableau de mesures est enregistré à différents instants.

1.3 Notion d'inertie

1.3.1 Nuage de points pondérés

Un tableau de données à n individus et p variables quantitatives définit un ensemble de n points dans $\mathbb{R}^p$ (espace des individus). On identifiera par commodité l'ensemble Ω des individus et l'ensemble des n points de $\mathbb{R}^p$:

$$\Omega = \{x_1, \ldots, x_n\} \text{ avec } x_i' = (x_i^1, \ldots, x_i^p) \in \mathbb{R}^p.$$

D'autre part, chacun de ces n points est muni de poids p_i strictement positifs et de somme généralement égale à 1. On dispose d'un nuage $N(\Omega)$ de n points pondérés dans $\mathbb{R}^p$:

$$N(\Omega) = \{(x_i, p_i), i = 1, \ldots, n\}$$

Au delà du cadre géométrique, l'analyse des données utilise les propriétés mécaniques d'un ensemble de points munis de poids et en particulier celles liées à l'inertie.

1.3.2 Inertie du nuage par rapport à un point

L'espace $\mathbb{R}^p$ étant muni d'une métrique euclidienne d_M (voir paragraphe 1.5.2) on appelle inertie de $N(\Omega)$ par rapport à un point a de $\mathbb{R}^p$ la quantité:

$$I_a = \sum_{i=1}^{n} p_i \, d_M^2(x_i, a)$$

1.3.3 Théorème de Huygens

Si

$$g = \sum_{i=1}^{n} p_i \, x_i$$

est le centre de gravité du nuage $N(\Omega)$ on a:

$$\forall a \in \mathbb{R}^p,\ I_a = I_g + \sum_{i=1}^{n} p_i\, d_M^2(a,g)$$

Remarques:

Le centre de gravité g est le point par rapport auquel l'inertie du nuage est minimum.

I_g est l'inertie totale du nuage $N(\Omega)$ et sera souvent notée I ou T.

1.3.4 Inerties associées à une partition

Soit $P = (P_1,\ldots,P_k)$ une partition en k classes de E. On note:

$$\mu_\ell = \sum_{i:\, x_i \in P_\ell} p_i$$

le poids de la classe P_ℓ, $\ell = 1, \ldots, k$.

On a bien sûr: $\sum_{\ell=1}^{k} \mu_\ell = 1$

et

$$g_\ell = \frac{1}{\mu_\ell} \sum_{x_i \in P_\ell} p_i\, x_i$$

est le centre de gravité de la classe P_ℓ, pour $\ell=1, \ldots, k$.

À cette partition P sont associées trois inerties:

L'inertie totale T, déjà rencontrée, indépendante de la partition:

$$T = \sum_{i=1}^{n} p_i\, d_M^2\,(x_i,g)$$

L'inertie interclasse B:

$$B = \sum_{\ell=1}^{k} \mu_\ell\, d_M^2\,(g_\ell,g)$$

Il s'agit de l'inertie du nuage des centres de gravité g_ℓ munis des poids μ_ℓ.

L'inertie intraclasse W:

$$W = \sum_{\ell=1}^{k} I_\ell$$

où I_ℓ est l'inertie de la classe P_ℓ par rapport à son propre centre de gravité g_ℓ:

$$I_\ell = \sum_{x_i \in P_\ell} p_i\, d_M^2\,(x_i,\, g_\ell).$$

L'inertie intraclasse est la moyenne des inerties des classes P_ℓ munies des poids μ_ℓ.

On a la relation fondamentale suivante:

$$\boxed{T = B + W}.$$

Cette relation se déduit directement du théorème de Huygens. Il résulte de cette formule que plus l'inertie intra-classe est faible plus l'inertie interclasse est grande.

Pour pouvoir analyser une partition, il est utile de décomposer les inerties associées suivant les classes et les variables. Nous présentons ces décompositions ici et nous fixons les notations liées à l'inertie. Dans tout ce paragraphe, nous supposons que la métrique M est diagonale ce qui est assuré la plupart du temps en analyse des données.

T peut s'écrire:

$$T = \sum_{\ell=1}^{k} \sum_{j=1}^{p} T_\ell^j$$

où T_ℓ^j est l'écart moyen, pour la variable j, des points de la classe P_ℓ au centre de gravité global, pour tout $\ell = 1, \ldots, k$; $j = 1, \ldots, p$.

B peut s'écrire:

$$B = \sum_{\ell=1}^{k} \sum_{j=1}^{p} B_\ell^j$$

où B_ℓ^j est l'écart, pour la variable j, du centre de gravité de la classe P_ℓ au centre de gravité global, pour tout $\ell = 1, \ldots, k$; $j = 1, \ldots, p$.

W peut s'écrire:

$$W = \sum_{\ell=1}^{k} \sum_{j=1}^{p} W_\ell^j$$

où W_ℓ^j est l'inertie de la classe P_ℓ pour la variable j, pour tout $\ell=1, \ldots, k$; $j = 1, \ldots, p$.

Par ailleurs, les différentes inerties se décomposent de manière additive sur les variables et sur les classes. En effet, si on note pour $j = 1, \ldots, p$

$$T^j = \sum_{\ell=1}^{k} T_\ell^j,\ B^j = \sum_{\ell=1}^{k} B_\ell^j \text{ et } W^j = \sum_{\ell=1}^{k} W_\ell^j,$$

on a

$$T = \sum_{j=1}^{p} T^j,\ B = \sum_{j=1}^{p} B^j \text{ et } W = \sum_{j=1}^{p} W^j.$$

T^j, B^j et W^j expriment la part de la variable j pour les inerties concernées. De même, on note pour tout ℓ=1, ..., k:

$$T_\ell = \sum_{j=1}^{p} T^j_\ell\ ,\ B_\ell = \sum_{j=1}^{p} B^j_\ell \text{ et } W_\ell = \sum_{j=1}^{p} W^j_\ell.$$

T_ℓ, B_ℓ et W_ℓ expriment la part de la classe P_ℓ pour les inerties concernées.

$$\text{On a } T = \sum_{\ell=1}^{k} T_\ell\ ,\ B = \sum_{\ell=1}^{k} B_\ell \text{ et } W = \sum_{\ell=1}^{k} W_\ell$$

Il est facile de voir que la relation fondamentale T = B + W se conserve par décomposition suivant j et k. On a:

$$\left.\begin{array}{l}\forall\ j = 1, \ldots, p \\ \forall\ \ell = 1, \ldots, k\end{array}\right\} \qquad T^j_\ell = B^j_\ell + W^j_\ell$$

$$\forall\ j = 1, \ldots, p \qquad T^j = B^j + W^j$$

$$\forall\ \ell = 1, \ldots, k \qquad T_\ell = B_\ell + W_\ell$$

Une relation importante

Étant donnés deux groupes disjoints P_1 et P_2 d'individus, on a la relation

$$I(P_1 \cup P_2) = I(P_1) + I(P_2) + \frac{\mu_1.\mu_2}{\mu_1+\mu_2}\, d^2_M\,(g_1, g_2)$$

où I(A) dénote l'inertie d'un groupe A.

Cette relation montre que la diminution d'inertie s'exprime simplement à l'aide des centres de gravité des classes.

1.3.5 Tableaux de contingence

Nuages de points associées

Étant donné un tableau de contingence défini sur les deux variables qualitatives I et J ayant respectivement n et p modalités, on peut définir deux nuages de points, le premier N(I) formé de n points dans l'espace $\mathbb{R}^p$ représente l'ensemble I, le second N(J) formé de p points dans $\mathbb{R}^n$ représente l'ensemble J.

Nuage N(I)

Afin de pouvoir comparer les modalités de I, on représente chaque modalité i par son profil f^i_j. Pour tenir compte de l'importance des modalités qui apparaissent plus fréquemment, on associe à chaque modalité i le poids f_i. (voir les notations définies en 1.2). On obtient ainsi un nuage de n points pesants dans $\mathbb{R}^p$, noté N(I).

$$N(I) = \{(f^i_j,\ f_{i.})\},\ i \in I\}$$

Ce nuage possède les propriétés suivantes:

- Le centre de gravité de N(I) est la loi marginale f_j.
- Le nuage N(I) est dans un sous-espace à p-1 dimensions de $\mathbb{R}^p$.

Nuage N(J)

On définit de la même façon un nuage N(J) de p points dans $\mathbb{R}^n$:

$$N(J) = \{(f^j_i,\ f_{.j}),\ j \in J\}$$

Cette fois, le centre de gravité du nuage est la loi marginale f_i et le nuage se situe dans un sous-espace à n-1 dimensions de $\mathbb{R}^n$.

Choix des métriques

Pour mesurer les proximités entre les points de ces deux nuages de profils, on utilise habituellement la métrique du χ^2.

La distance entre les modalités i et i' de I est définie par:

$$d^2\,(f^i_j,\ f^{i'}_j) = \sum_{j=1} (f^i_j - f^{i'}_j)^2 / f_{.j}$$

De même, pour $N(J) \subset \mathbb{R}^n$, la métrique du χ^2 est définie par:

$$d^2(f^j_i,\ f^{j'}_i) = \sum_{i=1}^{n} (f^j_i - f^{j'}_i)^2 / f_{i.}$$

Propriété d'équivalence distributionnelle

Si on remplace dans le tableau de contingence deux colonnes ayant le même profil, c'est-à-dire étant proportionnelles, par leur somme, on ne modifie pas les distances entre les profils du nuage N(I) et du nuage N(J). On a évidemment la même propriété d'invariance lorsque l'on regroupe deux lignes ayant le même profil.

Propriétés des inerties

Inertie des nuages N(I) *et* N(J).
L'inertie du nuage N(I) s'écrit:

$$T = \sum_{i \in I} f_{i.}\, d^2(f^i_j, f_j) = \sum_{i \in I} \sum_{j \in J} f_{i.}\ (f^i_j - f_j)^2 / f_{.j}$$
$$= \sum_{i \in I} \sum_{j \in J} (f_{ij} - f_{i.} f_{.j})^2 / (f_{i.} f_{.j})$$

En raison de la symétrie entre i et j, l'inertie du nuage N(J) est égale à l'inertie du nuage N(I).

Sachant que la statistique du χ^2 utilisée pour tester l'indépendance des variables I et J est

$$\chi^2(I,J) = \sum_{i \in I} \sum_{j \in J} \frac{(n_{ij} - s f_{i.} f_{.j})^2}{s f_{i.} f_{.j}}$$

(où $s = \sum_{i \in I} \sum_{j \in J} n_{ij}$), on obtient la relation

$$s\, I = \chi^2(I,J).$$

Cette quantité représente l'écart entre le produit des fréquences théoriques $f_{i.}\ f_{.j}$ que l'on aurait si il y avait indépendance entre les deux ensembles I et J et les fréquences observées f_{ij}. L'information apportée par le tableau sur les liens entre I et J sera d'autant plus grande que le χ^2 sera grand. Le χ^2 sera nul en cas d'indépendance entre I et J.

1.4 Changement de variable et codage

La question du changement de variable est fondamentale en analyse des données. Elle se pose au moins dans les trois cas suivants:

a) On a un tableau hétérogène et l'on désire exprimer certains des paramètres descriptifs à l'aide de nouvelles variables de façon que toutes les variables deviennent de même type.

b) Pour appliquer certaines méthodes d'analyse de données incompatibles avec le type des variables initiales, on sera amené par exemple, à transformer un tableau de données quantitatives en un tableau de modalités pour pouvoir utiliser, de façon efficace, une analyse factorielle des correspondances multiples.

c) Pour synthétiser l'information contenue dans un tableau de données, on réduira sa taille. On peut par exemple, remplacer l'ensemble des variables par quelques combinaisons linéaires de ces variables.

Il y a deux types de changement de variable: le changement de variable par changement de structure et le changement de variable par codage.

Changement de structure

Exemple: Considérons la variable tranche d'âge qui prend les valeurs 1 pour les jeunes, 2 pour les adultes, 3 pour les personnes âgées. Si l'on associe à cette variable la structure d'ordre usuelle, alors c'est une variable qualitative ordinale. Si maintenant, on ne désire plus faire intervenir l'ordre entre les tranches d'âge, la variable devient nominale. Dans une enquête sociologique, il peut être intéressant de supprimer l'ordre entre les tranches d'âge afin de faciliter l'apparition d'éventuels liens entre les jeunes et les personnes âgées.

Changement de codage

Exemple: Reprenons la variable v décrite dans l'exemple précédent. Si l'on désire mettre dans une même classe, les jeunes et les personnes âgées, on définit une variable v' qui prend, les valeurs: 1 et 2. v' est une variable qualitatitive nominale qui s'obtient par codage. Ce codage est tel que c(1) = c(3) = 1 et c(2) = 2.

Propriété d'équivalence distributionnelle

Si on remplace dans le tableau de contingence deux colonnes ayant le même profil, c'est-à-dire étant proportionnelles, par leur somme, on ne modifie pas les distances entre les profils du nuage N(I) et du nuage N(J). On a évidemment la même propriété d'invariance lorsque l'on regroupe deux lignes ayant le même profil.

Propriétés des inerties

Inertie des nuages N(I) *et* N(J).
L'inertie du nuage N(I) s'écrit:

$$T = \sum_{i \in I} f_{i.}\, d^2(f^i_j, f_j) = \sum_{i \in I} \sum_{j \in J} f_{i.}\ (f^i_j - f_j)^2 / f_{.j}$$
$$= \sum_{i \in I} \sum_{j \in J} (f_{ij} - f_{i.} f_{.j})^2 / (f_{i.} f_{.j})$$

En raison de la symétrie entre i et j, l'inertie du nuage N(J) est égale à l'inertie du nuage N(I).

Sachant que la statistique du χ^2 utilisée pour tester l'indépendance des variables I et J est

$$\chi^2(I,J) = \sum_{i \in I} \sum_{j \in J} \frac{(n_{ij} - s f_{i.} f_{.j})^2}{s f_{i.} f_{.j}}$$

(où $s = \sum_{i \in I} \sum_{j \in J} n_{ij}$), on obtient la relation

$$s\, I = \chi^2(I,J).$$

Cette quantité représente l'écart entre le produit des fréquences théoriques $f_{i.}\ f_{.j}$ que l'on aurait si il y avait indépendance entre les deux ensembles I et J et les fréquences observées f_{ij}. L'information apportée par le tableau sur les liens entre I et J sera d'autant plus grande que le χ^2 sera grand. Le χ^2 sera nul en cas d'indépendance entre I et J.

1.4 Changement de variable et codage

La question du changement de variable est fondamentale en analyse des données. Elle se pose au moins dans les trois cas suivants:

a) On a un tableau hétérogène et l'on désire exprimer certains des paramètres descriptifs à l'aide de nouvelles variables de façon que toutes les variables deviennent de même type.

b) Pour appliquer certaines méthodes d'analyse de données incompatibles avec le type des variables initiales, on sera amené par exemple, à transformer un tableau de données quantitatives en un tableau de modalités pour pouvoir utiliser, de façon efficace, une analyse factorielle des correspondances multiples.

c) Pour synthétiser l'information contenue dans un tableau de données, on réduira sa taille. On peut par exemple, remplacer l'ensemble des variables par quelques combinaisons linéaires de ces variables.

Il y a deux types de changement de variable: le changement de variable par changement de structure et le changement de variable par codage.

Changement de structure

Exemple: Considérons la variable tranche d'âge qui prend les valeurs 1 pour les jeunes, 2 pour les adultes, 3 pour les personnes âgées. Si l'on associe à cette variable la structure d'ordre usuelle, alors c'est une variable qualitative ordinale. Si maintenant, on ne désire plus faire intervenir l'ordre entre les tranches d'âge, la variable devient nominale. Dans une enquête sociologique, il peut être intéressant de supprimer l'ordre entre les tranches d'âge afin de faciliter l'apparition d'éventuels liens entre les jeunes et les personnes âgées.

Changement de codage

Exemple: Reprenons la variable v décrite dans l'exemple précédent. Si l'on désire mettre dans une même classe, les jeunes et les personnes âgées, on définit une variable v' qui prend, les valeurs: 1 et 2. v' est une variable qualitatitive nominale qui s'obtient par codage. Ce codage est tel que c(1) = c(3) = 1 et c(2) = 2.

Différents types de changement de variable

Transformation quantitatif-quantitatif:

Centrage réduction

Quand les variables sont mesurées avec des échelles différentes ou ont des dispersions hétérogènes, il peut s'avérer utile de "centrer" et "réduire" ces variables. Cela permet d'obtenir:

- des données indépendantes de l'échelle choisie,
- des variables ayant même moyenne et même dispersion.

. **Centrer** une variable v consiste à en soustraire sa moyenne.
. **Réduire** une variable v consiste à la diviser par son écart-type.

En centrant et réduisant une variable v, on obtient une variable v' qui est dite "centrée-réduite". Ainsi une variable v' centrée-réduite satisfait aux deux propriétés suivantes:

- sa moyenne est nulle
- son écart-type est égal à un.

Combinaison de variables

On peut créer une nouvelle variable quantitative en appliquant des fonctions numériques (combinaison linéaire, polynomiale, logarithme, etc).

On peut de plus utiliser des fonctions avec des expressions logiques provenant de variables quantitatives ou qualitatives.

Transformation quantitatif-qualitatif:

Découpage par bornes choisies par l'utilisateur

On effectue un découpage de $\mathbb{R}$ à l'aide de bornes définies par l'utilisateur. Un tel découpage étant réalisé, on numérote les classes associées à ce découpage en respectant l'ordre. La nouvelle variable qualitative ordinale est obtenue en affectant à chaque individu le numéro de la classe à laquelle il appartient. Remarquons qu'un tel découpage comme pour tous ceux que l'on verra par la suite entraîne une perte d'information de deux sortes. On perd:

- la distinction entre les objets d'une même classe,
- l'amplitude de la différence entre les objets de deux classes différentes.

Découpage par intervalles égaux

En utilisant les intervalles de même longueur I_1, ..., I_k, on obtient directement k classes d'objets contiguës. La longueur de chaque intervalle étant bien sûr égale à (M-m)/k où M et m sont respectivement la plus grande et la plus petite valeur prise par la variable.

Découpage par effectifs égaux

Ce type de découpage présente l'intérêt d'éviter les classes vides et découpera finement les endroits denses. Contrairement au découpage par intervalles égaux, il tient compte de l'échantillon.

Cette technique permet de trouver des classes d'individus ayant sensiblement le même effectif.

Découpage par minimisation de la variance

Ce découpage revient à chercher des classes contiguës P_1, ..., P_k dont l'inertie intraclasse est minimale.

Transformation qualitatif-qualitatif:

Par changement de structure ou par changement de codage.

Le premier type de transformation consiste à ne plus tenir compte de l'ordre d'une variable qualitative ordinale qui devient ainsi une variable qualitative nominale. Le deuxième type de transformation consiste à regrouper des modalités.

Par combinaison de variables.

Cette transformation est analogue à la combinaison logique de variables quantitatives.

Exemple: Soient deux variables qualitatives v_1 et v_2 à trois modalités codées 1, 2 et 3. On crée

une nouvelle variable qualitative v_3 à deux modalités codées 1 et 2:

$$v_3 = \begin{cases} 1 \text{ si } v_1 = 1 \text{ et si } v_2 = 1 \text{ ou } 3 \\ 2 \text{ sinon} \end{cases}$$

Transformation qualitatif-quantitatif:

Codage disjonctif complet

Ce codage consiste à transformer une variable qualitative à r modalités en r variables binaires indicatrices de chaque modalité.

Exemple: Soit trois individus a, b, c répondant aux questions suivantes: couleur des yeux (Y), âge (A), sexe (F), leurs réponses étant codées de la manière suivante:

Yeux	:	vert	: 1
		bleu	: 2
		marron	: 3
Âge	:	de 0 à 20 ans	: 1
		de 20 à 50 ans	: 2
		plus de 50 ans	: 3
Sexe	:	féminin	: 1
		masculin	: 2

Soit le tableau des réponses:

	Y	A	S
a	1	2	2
b	2	1	1
c	3	3	2

Le codage disjonctif complet de ce tableau est donné dans le tableau suivant :

	Yeux			Age			Sexe	
	vert	bleu	marron	0 à 20	20 à 50	+ de 50	F	M
a	1	0	0	0	1	0	0	1
b	0	1	0	1	0	0	1	0
c	0	0	1	0	0	1	0	1

Le codage est le suivant:
absence de la modalité: 0
présence de la modalité: 1

Codage additif.

Ce codage est de même nature que le codage disjonctif complet. Il n'a de sens que lorsque les variables à coder sont qualitatives ordinales. Il permet de conserver l'ordre entre les modalités. Par exemple, la variable âge du paragraphe précédent sera codée:

0 à 20 ans	:	(1, 0, 0)
20 à 50 ans	:	(1, 1, 0)
plus de 50 ans	:	(1, 1, 1).

1.5 Similarités et dissimilarités

1.5.1 Quelques définitions

Nous appelons similarité ou dissimilarité toute application à valeurs numériques qui permet de mesurer le lien entre les individus d'un même ensemble ou entre les variables. Pour une similarité le lien est d'autant plus fort que sa valeur est grande. Pour une dissimilarité le lien est d'autant plus fort que sa valeur est petite.

Définition d'un indice de similarité

Un indice de similarité (ou plus simplement une similarité) sur un ensemble Ω est une application s de $\Omega \times \Omega$ dans $\mathbb{R}^+$ qui vérifie les deux propriétés suivantes:

S1) s est symétrique: $\forall (w, w') \in \Omega \times \Omega$:
$$s(w, w') = s(w', w)$$
S2) $\forall (w, w') \in \Omega \times \Omega$ avec $w \neq w'$:
$$s(w, w) = s(w', w') > s(w, w')$$

Remarquons que la similarité d'un point à lui-même est une constante supérieure à tout autre similarité; on la note s_{max}.

Définition d'un indice de dissimilarité

Un indice de dissimilarité (ou plus simplement une dissimilarité) est une application d qui satisfait aux conditions données dans la définition d'un indice de similarité, sauf la deuxième qui est remplacée par $d(w, w) = 0$ pour tout $w \in \Omega$.

Définition d'une distance et d'une ultramétrique

Une distance est un indice de dissimilarité qui vérifie en plus les deux propriétés suivantes:

D1) $d(w,w') = 0 \Rightarrow w = w'$

D2) $d(w,w') \leq d(w, w'') + d(w'', w')$
$\forall(w,w',w'') \in \Omega^3$, (l'inégalité triangulaire).

Un indice de dissimilarité, qui vérifie seulement la propriété (D1), est appelé "**indice de distance**". S'il vérifie seulement la propriété (D2), on dit que c'est un "**écart**". Si au lieu de vérifier (D2), l'indice de dissimilarité vérifie l'inégalité suivante:

(D3) $d(w,w') \leq \text{Max}((d(w,w''), d(w'',w'))$
$\forall(w,w',w'') \in \Omega^3$.

On dit que c'est un "**écart ultramétrique**" (ou une "**dissimilarité ultramétrique**"). On voit facilement que la condition (D3) entraîne (D2).

Un indice de dissimilarité, satisfaisant à (D1) et (D3), est appelé "**distance ultramétrique**". Nous reviendrons dans le chapitre suivant sur cette distance qui joue un rôle important en classification hiérarchique.

Remarque: il est facile de transformer un indice de similarité s en un indice de dissimilarité s'. Il suffit de poser:

$$s'(w, w') = s_{max} - s(w, w').$$

1.5.2 Tableaux de variables quantitatives

Distances entre individus

Distances euclidiennes générales

Ce sont les distances les plus classiques.

Distance euclidienne simple:

$$d(w_i,w_{i'}) = \sqrt{\sum_{j=1}^{p} (x_i^j - x_{i'}^j)^2}$$

Distance de Mahalanobis: Elle se rencontre fréquement en analyse des données et surtout en analyse discriminante. Son expression analytique est la suivante:

$$d^2(w_i, w_{i'}) = {}^t(x_i - x_{i'})\ V^{-1}\ (x_i - x_{i'})$$

où V est la matrice de variance-covariance des individus.

Distance du χ^2: La distance du χ^2 est importante en analyse des données. Elle est particulièrement bien adaptée aux tableaux de contingence. Elle est utilisée en analyse factorielle des correspondances. Rappelons qu'elle s'exprime ainsi:

$$d^2(w_i, w_{i'}) = \sum_{j=1}^{p} \frac{1}{x_.^j} \left(\frac{x_i^j}{x_i^.} - \frac{x_{i'}^j}{x_{i'}^.} \right)^2$$

où $x_.^j = \sum_{i=1}^{n} x_i^j$ et $x_i^. = \sum_{j=1}^{p} x_i^j$

La distance de Minkowsky: Elle dépend d'un paramètre positif λ. On a:

$$d(w_i, w_{i'}) = \left[\sum_{j=1}^{p} \left| x_i^j - x_{i'}^j \right|^{\lambda} \right]^{1/\lambda}$$

Pour $\lambda = 2$, on retrouve la distance euclidienne simple, pour $\lambda = 1$, on a la distance des valeurs absolues:

$$d(w_i, w_{i'}) = \sum_{j=1}^{p} | x_i^j - x_{i'}^j |,$$

et finalement pour $\lambda \to +\infty$, on obtient la distance de Chebyshev:

$$d(w_i, w_{i'}) = \text{Max}_j | x_i^j - x_{i'}^j | .$$

Similarités entre variables

Les similarités les plus classiques sont la covariance ou corrélation entre variables. La valeur absolue de la corrélation est un indice de similarité.

1.5.3 Tableaux binaires

On peut, bien sûr, utiliser les distances vues précédemment. Si x_1 et x_2, sont deux variables binaires, pour définir une similarité ou une dissimilarité spécifique, il est nécessaire d'introduire les quatre quantités suivantes.

Soit a le nombre de fois où $x_1^j = x_2^j = 1$,

soit b le nombre de fois où $x_1^j = 0$ et $x_2^j = 1$,

soit c le nombre de fois où $x_1^j = 1$ et $x_2^j = 0$ et

soit d le nombre de fois où $x_1^j = x_2^j = 0$.

x_2 \ x_1	1	0
1	a	b
0	c	d

Les similarités suivantes sont souvent utilisées:

$$D_1(x_1, x_2) = \frac{a}{a+b+c+d} \quad \text{(Russel et Rao)}$$

$$D_2(x_1, x_2) = \frac{a}{a+b+c} \quad \text{(Jaccard)}$$

$$D_3(x_1, x_2) = \frac{2a}{2a+b+c} \quad \text{(Dice)}$$

$$D_4(x_1, x_2) = \frac{a}{a+2(b+c)} \quad \text{(Sokal et Sneath)}$$

$$D_5(x_1, x_2) = \frac{a+d}{a+b+c+d} \quad \text{(Sokal et Michener)}$$

$$D_6(x_1, x_2) = \frac{a}{a+b} + \frac{a}{a+c} \quad \text{(Kulzinsky)}$$

$$D_7(x_1, x_2) = \frac{a+d}{a+d+2(b+c)} \quad \text{(Rogers et Tanimoto)}$$

$$D_8(x_1, x_2) = \frac{ad-bc}{ad+bc} \quad \text{(Yule)}$$

$$D_9(x_1, x_2) = \frac{|\, ad-bc \,|}{[(a+b)\,(c+d)\,(a+c)\,(b+d)]^{1/2}}$$

(Corrélation, Pearson)

$$D_{10}(x_1, x_2) = \frac{a}{[(a+b)\,(c+d)\,(a+c)\,(b+d)]^{1/2}}$$

(Ochiai)

1.5.4 Tableaux de variables qualitatives

Similarités entre individus

Le codage disjonctif complet permet de se ramener à un tableau de variables binaires. On peut alors utiliser les similarités et dissimilarités indiquées précédemment et en particulier la distance du χ^2.

Similarités entre variables

À partir de deux variables qualitatives v_1 et v_2, on peut considérer le tableau de contingence associé. Ce tableau permet de définir une similarité entre les deux variables. Par exemple, la valeur du χ^2 de contingence peut être utilisée comme similarité entre les deux variables:

$$D(v_1, v_2) = \chi^2 = s \sum_{i=1}^{q} \sum_{j=1}^{r} \left(n_{ij} - \frac{n_i n_j}{s} \right)^2 \frac{1}{n_i n_j}$$

où q est le nombre de modalités de la variable v_1 et r le nombre de modalités de la variable v_2.

La "ressemblance" entre v_1 et v_2 est d'autant plus grande que $D(v_1, v_2)$ est grand, à effectif s fixé.

Si l'on désire une similarité liée au χ^2 et variant entre 0 et 1, on peut utiliser le T de Tschuprow:

$$D(v_1, v_2) = T = \left(\frac{\phi^2}{(r-1)\,(q-1)} \right)^{1/2} \quad \text{où } \phi^2 = \frac{\chi^2}{s}$$

Signalons encore quatre similarités basées sur le tableau de contingence:

$$D(v_1, v_2) = C = \left(\frac{\phi^2}{\min[(p-1),(q-1)]} \right)^{1/2}$$

(Coefficient C de Cramer).

Similarités utilisant l'information mutuelle de Shannon:

$$D(v_1, v_2) = I(v_1) + I(v_2) - I(v_1, v_2)$$

où
$$I(v_1) = \sum_{i=1}^{q} f_{i.} \log_2 f_{i.}$$

$$I(v_2) = \sum_{j=1}^{r} f_{.j} \log_2 f_{.j}$$

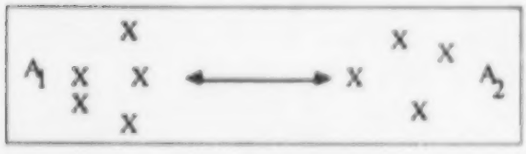
$$I(v_1, v_2) = \sum_{i,j} f_{ij} \log_2 f_{ij}$$

1.5.5 Dissimilarités entre groupes d'individus

Étant donné deux groupes d'individus A_1 et A_2 de Ω et une dissimilarité $\Omega \times \Omega \rightarrow \mathbb{R}^+$, on peut imaginer de nombreuses dissimilarités entre groupes. Citons en quelques-unes parmi les plus classiques

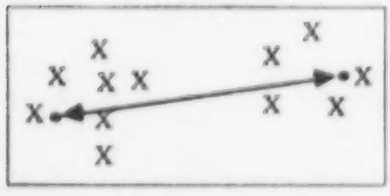
Distance du lien minimum:

$$D(A_1, A_2) = \min_{(w_i, w_j) \in A_1 \times A_2} d(w_i, w_j)$$



Distance du lien maximum:

$$D(A_1, A_2) = \max_{(w_i, w_j) \in A_1 \times A_2} d(w_i, w_j)$$



Distance des noyaux:

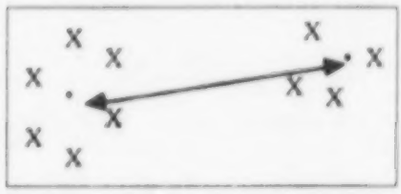
Un noyau N_j est formé de n_j éléments de A_j. La distance des noyaux est:

$$D(A_1, A_2) = \sum_{(w_i, w_j) \in N_1 \times N_2} d(w_i, w_j)$$

où N_j est le noyau de A_j.

Distance des centres de gravité:

Dans le cas où l'espace d'arrivée des p variables $O = O_1 \times \ldots \times O_p$ est un espace vectoriel, muni d'une distance quadratique, on peut prendre comme distance entre deux groupes, la distance de leur centre de gravité.



Dans le cas de deux groupes A_1 et A_2 disjoints, une distance dérivée de la distance des centres de gravité et plus souvent utilisée est la suivante:

$$D(A_1, A_2) = \frac{\mu(A_1)\,\mu(A_2)}{\mu(A_1)+\mu(A_2)}\, d^2(g_1, g_2)$$

où $\mu(A)$ est le poids du groupe A et où g_1 (resp. g_2) est le centre de gravité de A_1 (resp. A_2).

Cette distance représente la perte d'inertie en agrégant A_1 et A_2. On peut montrer effectivement que

$$D(A_1, A_2) = I(A_1, A_2) - I(A_1) - I(A_2)$$

où I(A) est l'inertie du groupe A.

2 Les méthodes de classifications automatiques

2.1 Introduction

La nature offre un grand nombre de populations qu'il est souhaitable de répartir en catégories. Chaque discipline scientifique sollicite des classifications. Il peut s'agir, en médecine par exemple, de découvrir les principaux regroupements de malades ayant le même comportement vis-à-vis de certains paramètres retenus pour les caractériser. En botanique, il s'agit de mettre en évidence des sous-espèces d'une même variété à partir d'un tableau de données où des plantes sont caractérisées par un certain nombre de

mensurations: les "taxons" d'où le terme de "taxonomie" utilisé parfois à la place de "classification automatique". En reconnaissance des formes, la classification automatique peut être utilisée pour obtenir des types d'empreintes digitales, d'écritures, d'éléctrocardiogrammes, de signaux radars. Plus généralement, on peut dire que la classification est à la base de toute connaissance. Le langage aussi est une forme de classification: chaque concept exprime une classe d'occurrence de faits de même type. Par extension, en intelligence artificielle, la classification automatique est considérée comme un procédé apprenant à l'ordinateur une information d'ordre sémantique qui n'était pas dans le tableau initial sous forme claire: on parle alors d'apprentissage sans professeur.

L'information apportée par une classification se situe, en effet, au niveau sémantique : "il ne s'agit pas d'atteindre un résultat vrai ou faux, probable ou improbable, mais seulement profitable ou non profitable" (LANCE et WILLIAMS, 1965). Profitable, pourquoi? Pour de multiples raisons: des regroupements inattendus apparaissent, permettant de faire de nouvelles hypothèses. Des regroupements attendus n'existent pas, ce qui fait ressortir le faible pouvoir séparateur des paramètres utilisés. Les classes obtenues et leurs imbrications assurent une vue concise et structurée des données. Les classes significatives entraînent la définition de fonctions de décision permettant d'attribuer un nouvel objet à la classe dont il est proche, etc.

Il existe maintenant de nombreux ouvrages en langue française entièrement ou en partie consacrés à la classification automatique (voir par exemple BENZÉCRI et al. (1980), DIDAY, LEMAIRE, POUGET, TESTU (1982), JAMBU, LEBEAUX (1978), LERMAN (1981), CAILLEZ, PAGES (1976)). Les principaux problèmes du domaine diffèrent suivant le type de classification recherchée: une partition, une hiérarchie, un recouvrement, ...

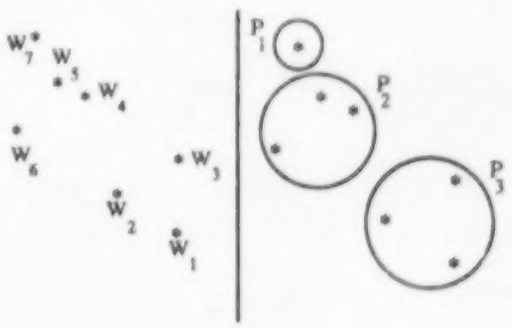
2.2 Les principaux espaces de classification

Pour mieux appréhender le contenu d'un tableau de données, on est souvent amené à rechercher des structures interclasses simples telles que les partitions et les recouvrements ou, plus complexes, telles que les hiérarchies ou les pyramides.

Les partitions

Une partition d'un ensemble Ω est un ensemble de parties non vides $P=(P_1,\ldots,P_k)$ d'intersections vides deux à deux et dont la réunion forme Ω.

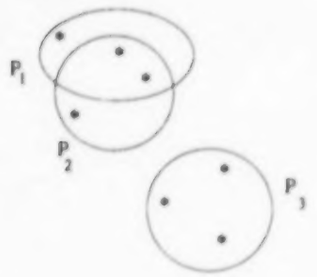
Exemple: Considérons la figure suivante. Étant donnés sept points du plan, on peut construire une partition en trois classes $P = (P_1, P_2, P_3)$, avec $P_1 = \{w_7\}$, $P_2 = \{w_5, w_4, w_6\}$ et $P_3 = \{w_1, w_2, w_3\}$.



Les recouvrements

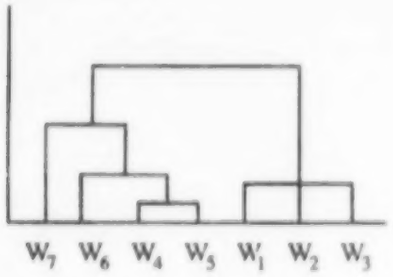
Un recouvrement de Ω est un ensemble $P = (P_1,\ldots,P_k)$ de parties non vides de Ω dont la réunion forme Ω. Une partition est donc un cas particulier de recouvrement.

Exemple: Étant donnés les sept points de la figure précédente, on peut construire le recouvrement à trois classes $P = (P_1,P_2,P_3)$ qui est représenté ci-dessus où $P_1 = \{w_7,w_5,w_4\}$, $P_2 = \{w_5,w_4,w_6\}$ et $P_3 = \{w_1,w_2,w_3\}$.



Les hiérarchies

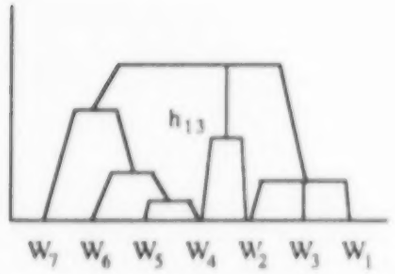
On cherche à représenter Ω par un ensemble de partitions emboîtées. En général, une hiérarchie est visualisée par un arbre appelé dendogramme. Par exemple, la hiérarchie de la figure suivante représente l'ensemble des sept points du plan déjà utilisés.



Exemple: Soit $\Omega = \{w_1,...,w_7\}$ les sept points considérés et H la hiérarchie qui leur est associée. On a $H = \bigcup_{i=1}^{12} h_i$ avec $h_i=\{w_i\}$ pour $i=1,7$; $h_8 = \{w_4, w_5\}$, $h_9 = \{w_1,w_2,w_3\}$, $h_{10}=\{w_6\} \cup h_8$, $h_{11}=\{w_7\} \cup h_{10}$ et $h_{12} = h_{11}\cup h_9$.

Les pyramides

Considérons encore une fois ces sept points. La représentation graphique de ces sept points telle qu'elle est indiquée est une "pyramide".



On voit que, contrairement aux hiérarchies, cette représentation autorise la possibilité pour un singleton ou un palier d'appartenir à deux classes. On remarque que si l'on coupe la pyramide par une suite de lignes horizontales, on obtient une suite de recouvrements emboîtés.

Exemple: Considérons la pyramide P précédente. Les paliers h_i de P sont les mêmes que ceux de la hiérarchie H de l'exemple précédent pour $i = 1, ..., 12$; on a donc $P = H\cup h_{13}$ avec $h_{13} = \{w_2,w_4\}$.

Formalisation de la notion d'espace de classification

Tous les ensembles de classes d'individus que nous avons présentés (partitions, recouvrements, hiérarchies, pyramides) satisfont à des propriétés communes. Considérons un ensemble Ω formé de x points et les classes qui leur sont associées. À partir de ces points, on peut définir plusieurs classifications "naturelles": si l'on considère que tous les points forment une seule classe, on obtient la classification définie par l'ensemble Ω lui-même; l'ensemble des classes $P_1, P_2, ..., P_k$ définit une deuxième classification; une troisième classification consiste à considérer que les points isolés forment x classes réduites chacune à un seul élément.

Considérons maintenant l'ensemble S formé de ces trois classifications. On a donc:

$$S=\{\ \{\Omega\},\{P_1, P_2, ..., P_k\},\ \{\{w_1\},\{w_2\},...,\{w_x\}\}\ \}$$

L'ensemble S satisfait aux quatre propriétés suivantes:

a) S est formé de trois éléments qui forment chacun une partie de l'ensemble des parties de Ω;

b) chacun de ces éléments est formé de classes dont la réunion recouvre Ω;

c) l'un des éléments de S contient Ω;

d) un autre contient les singletons.

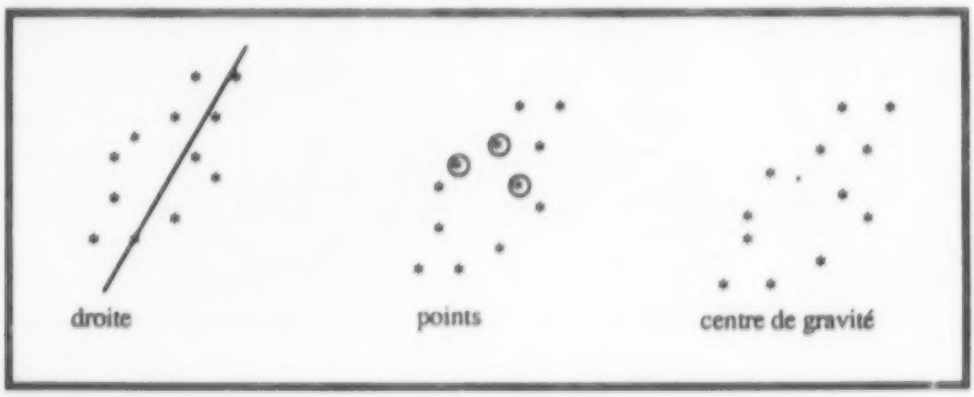
Ces quatre propriétés permettent de définir un espace de classification général qui englobe les différents types de classification que nous avons présentés.

Chaque élément $s \in S$ est une classification. La première condition signifie que chaque élément $s \in S$ est un ensemble de parties de Ω; la seconde, que tout élément de S recouvre Ω; la troisième qu'il existe au moins un élément de S contenant l'ensemble Ω et enfin la quatrième signifie que S contient un élément s qui contient lui-même tous les singletons. On a alors le résultat suivant: les partitions, les recouvrements, les hiérarchies et les pyramides sont des espaces de classification.

2.3 La méthode des nuées dynamiques

Dans cette section, nous montrons qu'une grande famille de problèmes de la classification automatique peuvent s'énoncer en termes d'optimisation d'un critère mathématiquement bien défini. La méthode des Nuées Dynamiques (appelée MND) largement développée dans DIDAY et al. (1980) fournit un cadre général permettant d'énoncer ces problèmes et d'obtenir des algorithmes pour leur donner des solutions approchées.

La caractéristique principale de la MND est la problématique nouvelle qu'elle a introduit. Il s'agit d'optimiser un critère qui exprime l'adéquation entre une classification des objets et un mode de représentation des classes de cette classification. Le problème d'optimisation se pose alors en terme de "recherche **simultanée** de la classification et de la représentation des classes de cette classification parmi un ensemble de classifications et de représentations possibles, qui optimisent le critère".



L'ensemble des partitions constitue un exemple privilégié d'espace de classification. On peut représenter chaque classe de la partition à l'aide d'un centre de gravité. L'ensemble de ces centres constitue un exemple classique d'ensemble de représentation. À partir des problèmes concrets posés par les utilisateurs, on peut

imaginer de nombreux autres espaces de représentation.

Les techniques de partitionnement dont nous allons parler maintenant ont pour but de fournir une partition en k classes d'individus (k donné a priori) bien agrégées et bien séparées entre elles. Elles présentent l'intérêt d'être rapides et de permettre le traitement de très grands tableaux (on a classifié, par exemple, 40 000 personnes ayant répondu à une enquête pour une entreprise de vente par correspondance afin d'obtenir des profils types de clientèle).

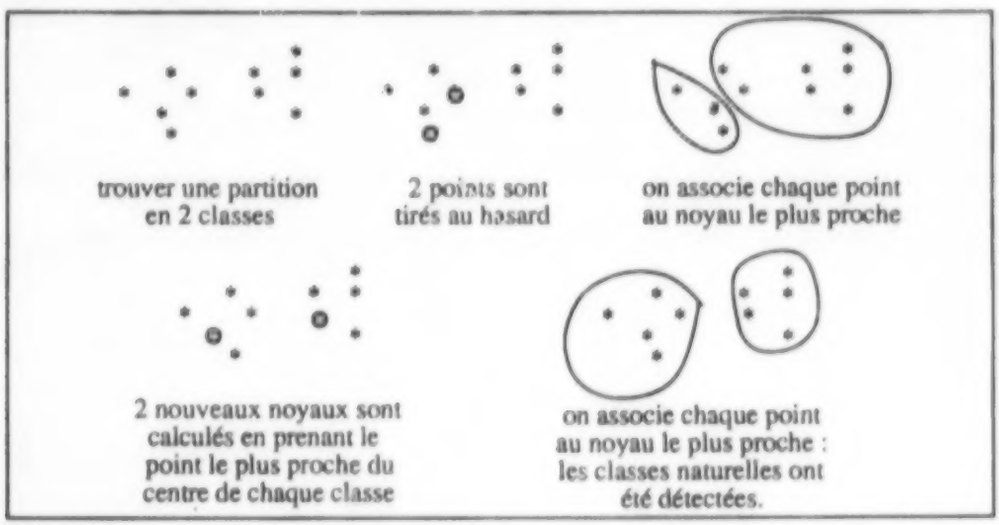
Choix d'un mode de représentation

Cet algorithme nécessite tout d'abord la définition d'un mode de représentation symbolique de tout groupe d'individus. Un groupe d'individus étant donné, cette représentation symbolique, appelée aussi "noyau", peut être, par exemple (figure précédante):

- une droite
- un groupe de points de la population,
- un centre de gravité.

Comment se déroule l'algorithme ?

Voyons un exemple dans le cas simple où le noyau est un point. On part d'un choix de k noyaux estimés ou tirés au hasard pris parmi une famille de noyaux admissibles appelée espace de représentation. Chaque point de la population est ensuite affecté au noyau dont il est le plus "proche". On obtient ainsi une partition en k



classes dont on calcule les noyaux. On recommence le procédé avec les nouveaux noyaux et ainsi de suite. On démontre que, sous certaines conditions, l'algorithme converge vers une position stable en améliorant, à chaque itération, un critère mathématique.

Deux nouveaux noyaux C et D sont calculés en prenant le point le plus proche du centre de chaque classe de l'étape précédente. On associe chaque point au noyau le plus proche, les classes naturelles ont donc été détectées par une procédé automatique.

Sous certaines conditions (portant sur les fonctions qui permettent d'affecter les points aux classes et de calculer les noyaux) cet algorithme fait décroître un critère W qui mesure l'adéquation entre les classes et leur noyau respectif (autrement dit la "ressemblance" des noyaux à leur classe). Ainsi, à chaque itération de l'algorithme, la décroissance du critère exprime une augmentation globale de l'adéquation entre les classes et leur noyaux.

Mode de représentation	Nom	Référence
centre de gravité	minimisation de la variance centres mobiles, K-means	(DIDAY 1988)
loi de probabilité	décomposition de mélanges	(DIDAY 1988)
distance	distance(s) adaptative(s)	(DIDAY 1988)
régression	régression typologique	(CHARLES 1977)
axe factoriel	analyse factorielle typologique	(SAKUN 1975)
axe discriminant	analyse discriminante typologique	(LEMOINE 1979)
variables	segmentation typologique	(MEUNIER 1986)
individus	nuées dynamiques sur tableaux de distances	(DIDAY 1988)
courbes	lissage typologique	(SAKUN 1975)

2.3.1 Tableau des principaux modes de représentation utilisés

Dans le tableau précédent qui n'est pas exhaustif on indique succinctement le mode de représentation, le nom et les références de la méthode qui en résulte.

On peut obtenir plusieurs solutions en changeant le choix de la partition initiale. Si une seule solution doit être retenue, on choisira naturellement celle qui donne la meilleure valeur du critère. Il est intéressant d'étudier l'ensemble des solutions afin d'en extraire des informations qui peuvent être utiles dans la pratique; par exemple, les groupes d'individus (appelés "formes fortes") qui se trouvent toujours dans la même classe quelle que soit la partition envisagée.

La méthode des Nuées Dynamiques comme la plupart des méthodes de partitionnement fournit une solution dépendant de la configuration initiale et nécessite le choix du nombre de classes. Les deux problèmes ne sont pas résolus. En général, le nombre de classes est demandé à l'utilisateur et l'initialisation est faite par tirage au hasard.

2.4 La Méthode des Nuées Dynamiques dans le cas des centres de gravité

Cette variante, appelée encore "algorithme des centres mobiles" (cf. BENZÉCRI 1973) est très utilisée car elle est efficace pour traiter de grands ensembles de données. Ce type d'algorithme, où une classe est représentée par son centre de gravité, a été étudiée par différents auteurs (THORNDIKE 1953, BONNER 1964, FORGY 1965, BALL ET HALL 1965 et MAC QUEEN 1967).

2.4.1 Les principales étapes

Dans cette partie, l'espace des individus est l'espace $\mathbf{R}^p$ muni d'une métrique euclidienne d_M. L'espace de représentation $\mathbf{L}$ d'une classe est aussi l'espace $\mathbf{R}^p$. La mesure d'adéquation D est une application de $P(\Omega) \times \mathbf{L}$ dans $\mathbf{R}^+$ définie de la manière suivante :

$$\forall A \in P(\Omega), \forall x \in \mathbf{R}^p,$$

$$D(A,x) = \sum_{a \in A} p_a \, d^2_M(x,a) = I_x(A)$$

où p_a est le poids de a et $I_x(A)$ l'inertie de la partie A par rapport à x.

Considérons le problème d'optimisation suivant:

$$\text{minimiser } D(A,x) = I_x(A) \text{ pour } x \in \mathbf{R}^p$$

D'après le théorème de Huygens, ce problème admet une solution unique qui est le centre de gravité de la partie A.

Le problème d'optimisation
Soit P_ℓ défini comme suit:

$$P_\ell = \{x \in \Omega \mid d(x,L_\ell) \leq d(x,L_m),$$

$$\forall m \in \{1, ..., K\} \text{ et } \ell < m \text{ en cas d'égalité}\}$$

Il s'agit de chercher le meilleur couple $(P,L) \in \mathbf{P}_k \times \mathbf{L}^k$ minimisant le critère d'adéquation W entre la partition $P = (P_1, ..., P_k)$ et son représentant $L = (L_1, ..., L_k)$. Le critère s'écrit:

$$W(P,L) = \sum_{\ell=1}^{k} D(P_\ell,L_\ell)$$

$$= \sum_{\ell=1}^{k} \sum_{x_i \in P_\ell} p_i \, d^2_M(x_i,L_\ell)$$

Comme le représentant L_ℓ d'une classe P_ℓ est son centre de gravité g_ℓ, le critère s'écrit encore:

$$W(P,L) = \sum_{\ell=1}^{k} \sum_{x_i \in P_\ell} p_i \, d^2_M(x_i,g_\ell) = \sum_{\ell=1}^{k} T_\ell$$

où T_ℓ est l'inertie de la classe P_ℓ par rapport a son centre de gravité g_ℓ. Le critère W(P,L) est donc l'inertie intraclasse de la partition P. Cette méthode minimise l'inertie intraclasse et maximise l'inertie interclasse en vertu de la relation: $T = W + B$.

2.4.2 Expressions du critère

Nous allons exprimer les deux aspects du critère W et B dans différents cas.

Les variables sont quantitatives
Deux cas importants sont envisagés:
- la métrique $M = I_p$ lorsque les données sont homogènes,
- la métrique $M = D_{1/\sigma^2}$ (les variables ont alors même variance) lorsque les données sont hétérogènes.

i) $M = I_p$
Notons μ_ℓ le poids de la classe ℓ, l'inertie interclasse B s'écrit:

$$B = \sum_{\ell=1}^{k} \mu_\ell d^2(g_\ell,g) = \sum_{\ell=1}^{k} \sum_{j=1}^{p} \mu_\ell (g^j_\ell - g^j)^2$$

où g^j_ℓ désigne la moyenne de la variable j dans la classe ℓ et g^j la moyenne de la variable j pour la population totale.

$$W = \sum_{\ell=1}^{k} \sum_{x_i \in P_\ell} p_i \, d^2(x_i,g_\ell)$$

$$= \sum_{\ell=1}^{k} \sum_{x_i \in P_\ell} \sum_{j=1}^{j} p_i (x^j_i - g^j_\ell)^2$$

$$= \sum_{\ell=1}^{k} \mu_\ell \, var_\ell(x^j)$$

si l'on désigne par $var_\ell(x^j)$ la variance locale de la variable x^j dans la classe ℓ.

ii) $M = D_{1/\sigma^2}$
On a alors:

$$B = \sum_{\ell=1}^{k} \mu_\ell \, d^2_{D_{1/\sigma^2}}(g_\ell, g)$$

$$B = \sum_{\ell=1}^{k} \sum_{j=1}^{p} \mu_\ell \left(\frac{g_\ell^j - g^j}{\sigma_j} \right)^2$$

$$W = \sum_{\ell=1}^{k} \sum_{x_i \in P_\ell} \sum_{j=1}^{p} p_i \left(\frac{x_i^j - g_\ell^j}{\sigma_j} \right)^2$$

Remarque: choisir $M = D_{1/\sigma^2}$ revient à traiter les données réduites avec $M = I_p$.

La méthode des centres mobiles maximise donc la somme des écarts entre les moyennes des variables dans les classes et les moyennes générales. En minimisant l'inertie intraclasse, la méthode a tendance à chercher des classes rondes, d'égal volume et de faibles inerties.

Les variables sont qualitatives

On considère un ensemble de p variables qualitatives $\{X_q: q \in Q\}$ où Q est l'ensemble des entiers allant de 1 à p. On a vu au chapitre 1 comment mettre un tableau de variables qualitatives sous forme disjonctive complète. Dans la suite, le tableau X désignera le tableau disjonctif complet associé aux variables qualitatives.

On notera:

- J_q l'ensemble des modalités de la variable X_q et card $X_q = m_q$.
- $X_q = \{X^j: j \in J_q\}$ la $j^{ème}$ modalité de X_q est représentée par le vecteur x^j de $\mathbf{R}^n$ tel que

$$x_i^j = \begin{cases} 1 \text{ si l'individu i a la modalité j} \\ 0 \text{ sinon} \end{cases}$$

- X_q est obtenue en juxtaposant les indicatrices des modalités de la variable.
- $N_q(\Omega)$ le nuage partiel des individus induit par la variable X_q.
- $J = \bigcup_q J_q$ l'ensemble total des modalités et

$$\text{card } J = \sum_q \text{card } J_q = m$$

(les ensembles J_q sont disjoints deux à deux).

- $X = (X_1, \ldots, X_p)$ le tableau de dimension $n \times m$ juxtaposant les tableaux X_q.

Tableau, nuage et centre de gravité:

Le tableau X ne comporte que des 0 et des 1, c'est un tableau de contingence particulier. Il est donc possible de reprendre le formalisme (profils, métrique du khi-deux,...) défini pour l'étude d'un tableau de contingence. Le tableau des profils des individus est égal à $X/\text{card}(Q)$. Étudier le tableau des profils revient à analyser le tableau X. On représentera un individu i, non par son profil, mais par le vecteur ligne x_i de X:

$$x_i \in \mathbf{R}^m, \sum_{j \in J} x_i^j = \text{card}(Q).$$

La somme des termes de chaque ligne et donc les fréquences marginales des lignes sont constantes. Les individus ont donc tous le même poids $p_i = 1/n$.

L'espace des individus $\mathbf{R}^m$ est muni d'une métrique euclidienne diagonale M quelconque. Les termes diagonaux de M sont: diag M = $\{q_j: j \in J\}$. Les coefficients q_j s'interprètent comme des pondérations associées aux modalités x^j intervenant notamment dans le calcul des distances entre les individus. La plupart du temps M sera la métrique du khi-deux.

Calculons le centre de gravité

$$g = \sum_i p_i x_i$$

du nuage des individus $N(\Omega) = \{(x_i, p_i): i \in I\}$. Le vecteur g a pour $j^{ème}$ composante:

$$g^j = \frac{1}{n} \sum_i x_i^j = \frac{n_j}{n} = f_j$$

où n_j et f_j sont respectivement l'effectif et la fréquence de la modalité j dans la population totale.

Expression de l'inertie interclasse B:

À une partition P ayant k classes, on associe une variable qualitative $Y = \{y_\ell: \ell = 1, \ldots, k\}$ ayant k modalités, y_ℓ est la variable indicatrice de la classe P_ℓ.

Le poids de la classe ℓ noté μ_ℓ a pour expression:

$$\mu_\ell = \sum_{x_i \in P_\ell} p_i = \frac{n_\ell}{n}$$

où n_ℓ est l'effectif de la classe P_ℓ. μ_ℓ s'interprète comme la fréquence de la classe P_ℓ dans la population totale.

Le centre de gravité de la classe P_ℓ est:

$$g_\ell = \sum_{x_i \in P_\ell} \frac{p_i}{\mu_\ell} x_i.$$

Ce vecteur a pour j-ème composante:

$$g_\ell^j = \sum_{x_i \in P_\ell} \frac{1/n}{n_\ell/n} x_i^j = \frac{n_{j\ell}}{n_\ell}$$

où $n_{j\ell}$ est l'effectif de la modalité j dans la classe ℓ. La composante g_ℓ^j est la fréquence de la modalité j dans la classe ℓ.

L'inertie interclasse B a pour expression:

$$B = \sum_{\ell=1}^{k} \mu_\ell \, d_M^2 (g_\ell, g) = \sum_{\ell=1}^{k} \sum_{j \in J} \mu_\ell \, q_j (g_\ell^j - g^j)^2$$

d'où

$$B = \sum_{q \in Q} \left(\sum_{\ell=1}^{k} \sum_{j \in J_q} \mu_\ell \, q_j (g_\ell^j - g^j)^2 \right)$$

On a donc:

$$B = \sum_{q \in Q} B_q$$

où

$$B_q = \sum_{\ell=1}^{k} \sum_{j \in J_q} \mu_\ell \, q_j \, (g_\ell^j - g^j)^2$$

est l'inertie interclasse de la partition P relative à la variable X_q. Remplaçons μ_ℓ, g^j et g_ℓ^j par leurs valeurs:

$$B_q = \sum_{\ell=1}^{k} \sum_{j \in J_q} \frac{n_\ell}{n} q_j \left(\frac{n_{j.\ell}}{n_\ell} - \frac{n_j}{n} \right)^2$$

$$= \sum_{\ell=1}^{k} \sum_{j \in J_q} q_j \frac{(n n_{j\ell} - n_j n_\ell)^2}{n^3 n_\ell}$$

Si M est la métrique du khi-deux c'est-à-dire, diag $M = \{q_j = \frac{1}{f_j} = \frac{n}{n_j} : j \in J\}$, on a:

$$B_q = \sum_{\ell=1}^{k} \sum_{j \in J_q} \frac{1}{n^2} \frac{(n_{j\ell} n - n_j n_\ell)^2}{n_j n_\ell} = \Phi_q^2$$

où Φ_q^2 est le phi-deux mesurant la liaison entre les variables qualitatives Y et X_q. L'inertie interclasse B a pour expression dans ce cas:

$$B = \sum_q \Phi_q^2$$

La méthode proposée, maximisant le critère B, recherche une variable qualitative "partition" Y la plus liée en moyenne au sens du phi-deux aux différentes variables X_q.

2.4.3 Méthodes voisines

La méthode proposée est intéressante car non seulement elle est efficace pour traiter des données importantes mais encore elle permet d'analyser des tableaux de variables de types divers. De nombreuses variantes existent. On peut citer la méthode des transferts (RÉGNIER 1965), les algorithmes à seuil où un individu n'est pas affecté à une classe si sa distance est supérieure à un seuil fixé par l'utilisateur. Par ailleurs d'autres méthodes cherchent une partition en affectant les objets aux centres des classes dont ils sont les plus proches et recommence le processus en découpant ou en regroupant les classes à l'aide de seuils suivant l'importance de la variance des classes. Tous ces algorithmes n'entrent pas dans le cadre des Nuées Dynamiques.

2.5 Classification avec distances adaptatives

2.5.1 Introduction

Dans tout problème de classification le choix de la mesure de ressemblance entre les objets à classer est fondamental. Lorsque les objets à classer sont décrits par p variables quantitatives, ils sont caractérisés par des vecteurs de $\mathbf{R}^p$ que

l'on munit généralement d'une métrique euclidienne:

$$\forall (x,y) \in \mathbb{R}^p \; d^2_M(x,y) = {}^t(x-y)\, M\, (x-y)$$

L'un des algorithmes de partitionnement le plus utilisé est l'algorithme des centres mobiles décrit dans le paragraphe précédent. Sa grande popularité tient à sa simplicité et à sa rapidité. Ainsi, il est particulièrement efficace pour classer de très grands tableaux de données (cf. LEBART, MORINEAU, et TABARD 1977). Cependant, on a constaté expérimentalement que ce choix de distance ($M=I_p$) a tendance à donner des classes sphériques de même volume, ce qui limite quelque peu ses champs d'application.

Pour lever cette difficulté, on peut prendre $M=T^{-1}$, T étant la matrice d'inertie des individus. Cela permet de reconnaître des classes de même allongement.

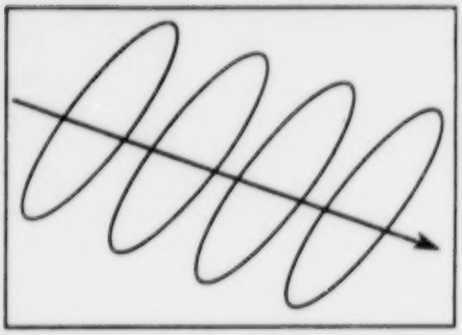
Certains auteurs se sont attachés à dépasser cette limitation de l'algorithme des centres mobiles (CHERNOFF 1970, FRIEDMAN et RUBIN 1967, GOVAERT 1975). Pour ce faire, ils ont supprimé l'astreinte pour la mesure de dissemblance d'être fixe tout au long de l'algorithme tout en restant dans le cadre quadratique de façon à obtenir une mesure de dissemblance qui s'adapte aux données traitées.

2.5.2 Une distance adaptative unique

Objectif

Le but de la méthode que nous présentons ici (FRIEDMAN et RUBIN 1967, GOVAERT 1975) est de reconnaître des classes ayant le même type de dispersion mais possédant des directions d'allongement inconnues.

Soit quatre classes à reconnaître qui sont assimilables à des ellipses de mêmes directions principales. Ces classes ne peuvent être reconnues par l'algorithme des centres mobiles pour peu qu'elles soient assez proches.



La solution proposée

Nous allons traiter ce problème en utilisant la méthode des Nuées Dynamiques. L'idée de base consiste à associer à une partition une métrique de telle sorte que la somme des distances de chaque individu de Ω au centre de gravité de sa classe soit la plus petite possible. Dans ce cadre une partition P de Ω sera représentée par k centres de gravité $g_1, \ldots, g_k$ et une métrique M de déterminant 1.

Le critère W est défini par

$$W(P,L) = \sum_{\ell=1}^{k} \sum_{x_i \in P_\ell} d^2_M (x_i, g_\ell)$$

où $P = (P_1, \ldots, P_k)$ et $L = (g_1, \ldots, g_k, M)$.

Ce critère représente la somme des inerties de chaque classe par rapport à son centre de gravité au sens de la distance d_M, chaque individu étant muni du poids 1.

La partition associée à $L = (g_1, \ldots, g_k, M)$ et minimise $W(\cdot,L)$ est obtenue ainsi:
$\forall \ell = 1, \ldots, k$

$$P_\ell = \{x_i \in \Omega : d_M(x_i,g_\ell) < d_M(x_i,g_m)\ \forall m$$
$$\text{avec } \ell < m \text{ en cas d'égalité}\}$$

La représentation $L = (g_1, \ldots, g_k, M)$ est associée à $P = (P_1, \ldots, P_k)$ de manière à minimiser le critère $W(P, \cdot)$. Pour $\ell = 1, \ldots, k$, g_ℓ est le centre de gravité de la classe P_ℓ.

La détermination de M est plus complexe: on cherche $M \in \mathbf{M}$ telle que

$$\sum_{\ell=1}^{k} \sum_{x_i \in P_\ell} {}^t(x_i - g_\ell) M (x_i - g_\ell) \text{ soit minimum.}$$

Il s'agit du même problème que celui rencontré en analyse discriminante linéaire pour choisir la métrique la plus discriminante (voir GOVAERT 1975). La solution est fournie par la matrice M définie par:

$$M = |W|^{1/p} W^{-1}$$

où W est la matrice d'inertie intraclasse associée a la partition P. L'algorithme des Nuées Dynamiques ainsi défini fait décroître le critère à chaque itération.

2.5.3 Une distance adaptative par classe

Objectif

L'algorithme précédent ne permet pas de reconnaître des classes présentant des directions d'allongement différentes et inconnues.

Le but de l'algorithme des Nuées Dynamiques que nous présentons maintenant est précisément de pouvoir reconnaître ce type de classes.

La solution proposée

L'idée de base consiste à associer à chaque classe de la partition une métrique de telle sorte que la somme des distances de chaque individu de Ω au centre de gravité de sa classe au sens de la métrique associée à sa classe soit la plus petite possible. Chacune des métriques (M_ℓ, ℓ=1, ..., k) appartient à M. Dans ce cadre, une partition P de Ω sera représentée par k centres de gravités g_1, ..., g_k et k matrices M_1, ..., M_k appartenant à $\mathbf{M}$. L'espace de représentation est donc: $L_k = (\mathbb{R}^p)^k \times (\mathbf{M})^k$.

Le critère W est défini par:

$$W(P,L) = \sum_{\ell=1}^{k} \sum_{x_i \in P_\ell} d^2_{M_\ell}(x_i, g_\ell)$$

où

$$P = (P_1,...,P_k) \text{ et } L = (g_1,...,g_k, M_1,...,M_k)$$

Ce critère représente la somme des inerties de chaque classe par rapport à son centre de gravité au sens de la distance d_{M_ℓ} qui lui est associée.

La partition associée à $L = (g_1,...,g_k, M_1, ..., M_k)$ minimisant $W(\cdot,L)$ est obtenue ainsi:

$\forall u = 1, ..., k$

$$P_\ell = \{x_i \in \Omega \,/\, d_{M_\ell}(x_i,g_\ell) \le d_{M_m}(x_i,g_m)$$
$$\forall m \text{ avec } \ell < m \text{ en cas d'égalité}\}$$

La représentation $L = (g_1, ..., g_k, M_1, ..., M_k)$ est associée à $P = (P_1,...,P_k)$ de manière à minimiser le critère $W(P,\cdot)$. On peut voir que $\forall \ell = 1, ..., k$ g_ℓ est le centre de gravité de la classe P_ℓ. Quant aux matrices M_ℓ ($\ell = 1, ...,$ k) elles doivent minimiser

$$\sum_{x_i \in P_\ell} {}^t(x_i - g_\ell) M_\ell (x_i - g_\ell),\ M_\ell \in \mathbf{M}.$$

Les solutions sont fournies par les matrices (M_ℓ, ℓ=1, ..., k) définies par

$$\forall \ell = 1, ..., k,\ M_\ell = |W_\ell|^{1/p} W_\ell^{-1}$$

où

$$W_\ell = \sum_{x_i \in P_\ell} (x_i - g_\ell)\, {}^t(x_i - g_\ell)$$

est la matrice d'inertie de la classe P_ℓ.

L'algorithme des Nuées Dynamiques ainsi défini fait décroître le critère à chaque itération.

2.5.4 Considérations pratiques sur les deux méthodes

Une analyse superficielle pourrait faire croire que l'algorithme des distances adaptatives est supérieur à l'algorithme des centres mobiles. La réalité est différente.

D'une part, les méthodes avec distances adaptatives imposent l'inversion de matrices de taille (p×p) à chaque itération. Elles sont donc très coûteuses (surtout la seconde).

D'autre part, à chaque itération le nombre de paramètres à estimer est très important. Aussi

pour obtenir des résultats fiables, le nombre d'individus à analyser doit être grand relativement à p et k. Pour la même raison, ces algorithmes donnent des résultats dépendant beaucoup plus de l'initialisation que des algorithmes plus simples (comme l'algorithme des centres mobiles). On pallie en grande partie à cet inconvénient en commençant par quelques itérations de l'algorithme des centres mobiles.

Toutes ces réserves font que les algorithmes des distances adaptatives sont à utiliser avec précaution. En particulier, il faut disposer d'échantillons de taille respectable et s'assurer qu'un algorithme plus simple ne donne pas des résultats satisfaisants.

2.6 La classification croisée

2.6.1 Introduction

Dans le cas où les données mettent en jeu deux ensembles, ce qui est très fréquent, les méthodes de classification automatique, en ne faisant porter la structure recherchée que sur un seul des deux ensembles, agissent de facon dissymétrique et privilégient un des deux ensembles, contrairement aux méthodes factorielles qui fournissent simultanément des résultats sur les deux ensembles. Dans ce paragraphe, nous étudions des méthodes de classification qui fournissent simultanément une partition sur chacun des deux ensembles. Plusieurs types de tableaux sont envisagés: les tableaux de contingence, les tableaux de variables quantitatives, les tableaux binaires, les tableaux de variables qualitatives. Ces méthodes peuvent être rapprochées des méthodes de classification directe (HARTIGAN 1975), qui essaient d'organiser directement le tableau initial ou des méthodes manuelles de BERTIN (1977) dont le principe est de simplifier le tableau en permutant lignes et colonnes.

2.6.2 Un exemple

	1	2	3	4	5	6	7	8	9	10
a	1	0	1	0	1	0	0	1	0	1
b	0	1	0	1	0	1	1	0	1	0
c	1	0	0	0	0	0	0	1	1	0
d	1	0	1	0	1	0	0	1	0	1
e	0	1	0	1	0	1	1	0	1	0
f	0	1	0	1	0	1	1	0	1	0
g	0	1	0	0	0	0	0	1	0	1
h	1	0	1	1	0	0	0	1	1	0
i	1	0	0	1	0	0	0	0	0	1
j	0	1	0	1	0	1	1	0	0	0

Considérons le tableau précédent. L'ensemble des lignes correspond à un ensemble de 11 micro-ordinateurs et l'ensemble des colonnes à 10 propriétés que ces micro-ordinateurs peuvent posséder (valeur 1) ou ne pas posséder (valeur 0). On cherche à résumer ce tableau de la manière suivante: classer les deux ensembles de telle sorte que les classes de micro-ordinateurs soient bien caractérisées par les classes de propriétés et vice-versa, que les classes de propriétés soient bien caractérisées par les classes de micro-ordinateurs.

Dans notre exemple, si on considère les classifications $\{\{a, d, h\}, \{b, e, f, j\}, \{c, g, i\}\}$ et $\{\{1, 3, 5, 8, 10\}, \{2, 4, 6, 7, 9\}\}$ on obtient en réordonnant les lignes et les colonnes suivant ces deux partitions le tableau suivant:

		1					2				
		1	3	5	8	10	2	4	6	7	9
A	a	1	1	1	1	1	0	0	0	0	0
	d	1	1	0	1	0	0	0	0	0	0
	h	1	1	1	1	1	0	0	1	0	1
B	b	0	0	0	0	0	1	1	1	1	1
	e	0	0	0	0	0	1	1	1	1	1
	f	0	0	0	0	0	1	0	1	1	1
	j	0	0	0	0	0	1	1	0	1	0
C	c	1	0	0	1	0	0	0	0	0	1
	g	0	0	0	1	1	1	0	0	0	0
	i	1	0	0	0	1	0	1	0	0	0

Le tableau initial peut alors être résumé par le tableau croisant les 2 partitions:

	1	2
A	1	0
B	0	1
C	0	0

A, B et C correspondent aux 3 classes de micro-ordinateurs, 1 et 2 aux 2 classes de propriétés. On a associé à chaque case la valeur 0 ou 1 qui était majoritaire. Ce tableau permet de voir que les micro-ordinateurs de la classe A possède en général les propriétés 1 mais pas 2, ceux de la classe B les propriétés 2 mais pas 1 et ceux de la classe C aucune propriété. On a résumé ainsi le tableau de 100 valeurs par un tableau de 6 valeurs. Le premier problème sera bien sûr de définir un critère numérique mesurant la qualité du résumé. Il reste ensuite à trouver le couple de partitions optimisant ce critère. Pour répondre à ces problèmes, une famille d'algorithmes de classification croisée adaptés à la plupart des tableaux de données que l'on peut rencontrer a été développée pour les tableaux binaires (logiciel SICLA).

2.6.3 Principe général

Avant d'entrer dans le détail de ces quatre algorithmes, on donne ici le principe général de leur construction. Si on note I et J les deux ensembles sur lesquels est défini le tableau de données, on considère S l'ensemble de tous les couples de partitions de I et de J. À chaque fois, une fonction W: $S \longrightarrow \mathbf{R}^+$, appelé critère, mesurant la qualité d'un couple de partitions et dépendant du type de tableau de données est définie. Le problème est alors de déterminer l'élément S optimisant ce critère, c'est-à-dire le couple de partitions de I et J qui minimisera ou maximisera suivant le cas la fonction W. L'algorithme proposé pour résoudre ce problème construit une suite (P^n,Q^n) de couple de partitions. Cette suite est définie à partir d'un couple initial (P^o,Q^o) en appliquant le mécanisme suivant: on fixe une des partitions et on cherche alors la meilleure partition de l'autre ensemble, puis on fixe cette fois la partition que l'on vient de trouver et l'on cherche la meilleure partition du premier ensemble. On recommence ces 2 étapes jusqu'à la convergence. On obtient ainsi un algorithme itératif. Pour trouver à chaque étape la meilleure partition, on fera appel à une méthode des Nuées Dynamiques. Cette méthode étant elle aussi itérative, on aura deux niveaux d'itérations dans les algorithmes proposés. On obtiendra à chaque fois un algorithme qui convergera vers un optimum local.

2.6.4 Tableau de contingence

Introduction

L'algorithme permet d'analyser des tableaux de contingence définis sur deux ensembles I et J ou, de manière plus générale, des tableaux possèdant des propriétés équivalentes. Cette analyse, qui repose sur l'optimisation du critère du χ^2 de contingence, s'applique, comme l'analyse des correspondances, aux mêmes types de données. Disposant d'un tableau de contingence défini sur deux ensembles I et J, il s'agit de trouver une partition P de I en K classes et une partition Q de J en L classes telles que le χ^2 de contingence du nouveau tableau de contingence construit en regroupant les lignes et les colonnes suivant les partitions P et Q soit maximum. L'algorithme proposé construit une suite de couples de partitions (P^n,Q^n) optimisant le χ^2 du tableau de contingence en appliquant alternativement sur I et sur J une variante de la méthode des Nuées Dynamiques.

Notations

On reprendra les notations définies au chapitre 1 sur les tableaux de contingence. On notera $T_0 = (t_{ij})$ le tableau de contingence initial défini sur les 2 variables qualitatives I et J ayant respectivement n et p modalités. Rappelons que l'on peut définir la dépendance entre I et J par le χ^2 de contingence:

$$\chi^2(I,J) = s \sum_{i \in I} \sum_{j \in J} (f_{ij}-f_i.f_{.j})^2/(f_i.f_{.j})$$

Cette quantité représente l'écart entre les fréquences théoriques $f_i.f_{.j}$ que l'on aurait, s'il y avait indépendance entre les deux ensembles I et

J, et les fréquences observées f_{ij} Le χ^2 sera nul en cas d'indépendance entre I et J. Le χ^2 sera grand lorsque les liens entre I et J seront forts. Ce χ^2 nous servira à mesurer l'information apportée par un tableau de contingence.

χ^2 de contingence associé à un couple de partitions

À partir de deux partitions P et Q, on peut définir un nouveau tableau de contingence. Si $P = (P_1, ..., P_K)$ est la partition de I en K classes et $Q = (Q_1, ..., Q_L)$ est la partition de J en L classes, on définit le tableau de contingence $T_1(P,Q)$:

$$T_1(k,\ell) = \sum_{i \in P_k} \sum_{j \in Q_\ell} t_{ij}$$

$\forall\, k = 1, ..., K$ et $\forall\, \ell = 1, ..., L$. (On a fait la somme des éléments de chaque couple de classes P_k, Q_ℓ).

Le nouveau tableau de contingence est donc un tableau à K lignes et L colonnes. Comme pour le tableau T_0, on peut lui associer son χ^2 de contingence qui mesurera l'information apportée par le tableau T_1. On le notera $\chi^2(P,Q)$.

Ce χ^2 va être utilisé pour mesurer la qualité des deux partitions P et Q des ensembles I et J.

Le problème d'optimisation

On montrera plus loin que l'on a nécessairement

$$\chi^2(I,J) \geq \chi^2(P,Q).$$

Le regroupement des éléments de chaque classe conduit donc à une perte d'information. Chercher à minimiser cette perte d'information revient à chercher les partitions P et Q maximisant le χ^2 de contingence du tableau qui leur est associé, rendant ainsi maximum la dépendance entre la partition P et la partition Q. Le problème que l'on va donc essayer de résoudre est la recherche simultanée de deux partitions P et Q maximisant le χ^2 de contingence du tableau associé. Pour justifier l'intérêt de ce problème, reprenons l'exemple de BENZÉCRI (1973). Le tableau de contingence est défini sur un ensemble I de communes et sur un ensemble J de professions. t_{ij} représente le nombre de personnes pratiquant la profession j dans la commune i. "Une classification des communes par une partition P sera d'autant meilleure que la connaissance de la classe d'une commune i nous apportera plus d'information sur la part des différentes professions j dans les activités de cette commune. De même une classification Q de l'ensemble des professions qui se pratiquent sera d'autant meilleure qu'elle groupera ensemble les professions qui se pratiquent le plus dans les mêmes communes: ainsi la connaissance de la classe d'une profession j nous fera approximativement connaître la répartition entre les communes de ceux qui pratiquent j". Signalons que, dans le cas idéal où les profils en ligne et en colonne sont égaux à l'intérieur de chaque classe de P et Q, il n'y a pas de perte d'information.

Recherche de la partition optimale à chaque étape

L'algorithme recherche alternativement une partition de I et une partition de J. Pour obtenir ces partitions, on fait appel à chaque étape à une variante des Nuées Dynamiques que nous précisons maintenant. On n'envisage ici que la recherche d'une partition de I, les résultats seront bien évidemment semblables pour la recherche d'une partition de J. Pour trouver une partition de I en K classes, nous allons appliquer l'algorithme des Nuées Dynamiques avec comme noyaux les centres de gravité, sur le nuage N(I) des profils f^i_j munis des poids $f_{i.}$, avec la métrique du χ^2 de matrice D_{1/f_J}.

On obtient alors les résultats suivants:

a) le critère optimisé par cette méthode est

$$W(P) = \sum_{k=1}^{K} \sum_{i \in P_k} f_{i.}\ d^2(i,G(P_k))$$

où $P = (P_1, ... P_K)$, d est la distance D_{1/f_J} et $G(P_k)$ est le centre de gravité de P_k, c'est-à-dire:

$$G(P_k) = \{ \sum_{i \in P_k} f_{i.}(f_{ij}/f_{i.}) \} / \sum_{i \in P_k} f_{i.} = \sum_{i \in P_k} f_{ij} / \sum_{i \in P_k} f_{i.}$$

b) la suite (P^n) construite par l'algorithme est stationnaire et de plus

$$W(P^0) > W(P^1) > \ldots > W(P^N) = W(P^{N+1}) = \ldots$$

Lien entre le critère W et le χ^2 de contingence

Dans le chapitre 1, on a montré que:

$$\chi^2(I,J) = sW(P) + \chi^2(P,J),$$

sW(P) représente l'information perdue en effectuant les regroupements de la partition P, tandis que $\chi^2(P,J)$ correspond à l'information conservée. En conséquence, puisque la quantité $\chi^2(I,J)$ ne dépend pas de la partition P, la recherche de la partition minimisant le critère W(P) est équivalente à la recherche de la partition P maximisant $\chi^2(P,J)$. La méthode des Nuées Dynamiques décrite précédemment maximise donc le χ^2 de contingence du tableau (P,J). Par raison de symétrie, on obtient le même résultat lorsque l'on fait une partition sur l'autre ensemble:

$$\chi^2(P,Q) \leq \chi^2(I,J).$$

Indices d'interprétation d'un couple de partitions

Pour définir des indices spécifiques à la classification croisée, il est intéressant de considérer que l'ensemble des valeurs $f_{ij}/f_{i.}f_{.j}$ munies des poids $f_{i.}f_{.j}$ forme un nuage $N(I\times J)$ de $\mathbf{R}^p$. Le centre de gravité de ce nuage est alors la valeur 1. Si on munit de plus $\mathbf{R}$ de la métrique euclidienne identité, le χ^2 du tableau initial représente l'inertie de ce nuage. À chaque couple de classes (P_k,Q_ℓ) on peut associer la classe de $N(I\times J)$, notée (k,ℓ) formée par les valeurs f_{ij} obtenues lorsque i et j décrivent les classe P_k et Q_ℓ. On peut alors définir $T_{k\ell}$, inertie de la classe (k,ℓ) par rapport au centre de gravité de $N(I\times J)$, soit:

$$T_{k\ell} = \sum_{i\in P_k} \sum_{j\in Q_\ell} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

$T_{k\ell}$ représente donc la part apportée par les couples (i,j) de la classe (k,ℓ) à la valeur du χ^2 du tableau initial.

On définit l'inertie conservée par le centre de gravité de la classe (k,ℓ):

$$B_{k\ell} = \frac{(f_{k\ell} - f_{k.}f_{.\ell})^2}{f_{k.}f_{.\ell}}$$

$W_{k\ell}$ sera l'inertie de la classe (k,ℓ).

On a par le théorème de Huygens

$$T_{k\ell} = B_{k\ell} + W_{k\ell}$$

En outre, $\chi^2(I,J) = T = \sum_{k,\ell} T_{k\ell}$

et

$$\chi^2(P,Q) = B = \sum_{k,\ell} B_{k\ell}$$

Ces relations permettent de définir:

- le tableau des valeurs $B_{k\ell}/B$ représentant la part d'inertie conservée par classe. Ces quantités indiquent l'importance de la classe.
- le tableau des valeurs $B_{k\ell}/T_{k\ell}$ représentant la part d'inertie conservée par la classe par rapport à l'inertie initiale des points de la classe. Cette quantité indique la qualité de représentation d'une classe. La valeur obtenue comprise entre 0 et 1 sera d'autant plus grande que la classe sera homogène.

L'ensemble de ces indices a été construit grâce au croisement des deux partitions et de ce fait donne une interprétation plus fine que les indices habituels qu'il est possible d'appliquer sur les deux partitions séparément.

2.6.5 Tableaux de variables qualitatives

Comme pour l'analyse des correspondances multiples (LEBART, MORINEAU, TABARD 1977), on peut considérer le questionnaire, après l'avoir transformé en un tableau disjonctif complet comme un tableau de contingence. On peut alors lui appliquer la méthode précédente qui permet la recherche simultanée d'une partition des individus et d'une partition des modalités. L'inconvénient de cette approche est la nécessité de disposer du tableau disjonctif complet dont la taille peut vite devenir trop grande.

2.6.6 Tableaux de variables quantitatives

On s'intéresse maintenant aux tableaux de données définies sur un ensemble d'individus et sur un ensemble de variables quantitatives. Contrairement aux tableaux de contingence sur lesquels portaient les paragraphes précédents, les

tableaux de mesures présentent une dissymétrie. Les lignes et les colonnes ne sont pas traitées de la même façon. Les lignes sont formées d'unités ayant un caractère répétitif. Ce sont les individus. Les colonnes peuvent être plus hétérogènes. Ce sont les variables. Ces variables correspondent à des mesures. Comme pour l'Analyse des Composantes Principales, l'objet de ce paragraphe sera d'extraire l'essentiel de l'information contenue dans le tableau de données, les moyens utilisés étant, non pas la représentation sur des plans factoriels, mais le regroupement des lignes et des colonnes en classes homogènes. Pour cela, il existe un algorithme de classification simultanée des deux ensembles adapté à ce type de tableau (progiciel SICLA). Le critère optimisé est un critère d'inertie. Après avoir précisé les notations utilisées, nous munissons les ensembles d'individus et de variables de métriques permettant de mesurer l'information apportée par le tableau de données. À un couple de partitions des deux ensembles est alors associé un nouveau tableau de données. L'objectif de l'algorithme de classification croisée est la recherche d'un couple de partitions tel que la "perte d'information" due au regroupement soit minimum, c'est-à-dire tel que la différence entre l'information apportée par le tableau initial et celle apportée par le tableau obtenu après regroupement soit minimum. Les différences essentielles avec l'algorithme précédent concernent le choix de la mesure d'information et la définition du tableau associé à un couple de partitions. Les différences et les ressemblances entre les deux méthodes sont exactement les mêmes que celles qui existent entre l'Analyse en Composantes Principales et l'Analyse des Correspondances.

Notations

Soit Ω un ensemble de n individus décrits à l'aide d'un ensemble V de p variables ou caractères quantitatifs. Les données sont rangées dans un tableau de description X à n lignes et p colonnes:

$$\mathbf{X} = (x_i^j),\ i \in \Omega \text{ et } j \in V$$

où x_i^j est la valeur de la variable j pour l'individu i.

D'autre part, on supposera que les individus sont munis de poids p_i vérifiant

$$\sum_{i \in \Omega} p_i = 1$$

Si aucune indication contraire est fournie, ces poids seront égaux, on aura donc

$$p_i = \frac{1}{n},\ \forall i \in \Omega$$

Associons à chaque individu i le vecteur $x_i = (x_i^1, \ldots, x_i^p)$ de $\mathbf{R}^p$ correspondant à la ligne i du tableau X. L'ensemble des x_i munis des pondérations p_i forme un nuage $N(\Omega)$ contenu dans $\mathbf{R}^p$. Nous supposerons que le tableau X est centré en colonne, c'est-à-dire que la somme des éléments de chaque colonne est nulle. Si ce n'est pas le cas, il suffit d'une transformation simple pour satisfaire cette condition. Cette condition se traduit pour le nuage $N(\Omega)$ par le fait que son centre de gravité est à l'origine de l'espace $\mathbf{R}^p$

$$\sum_{i \in \Omega} p_i\, x_i^j = 0,\ \forall j \in V$$

De la même façon, à chaque variable j est associé le vecteur $x^j=(x_1^j, \ldots, x_n^j)$ correspondant à la colonne j du tableau X. L'ensemble des x^j munis de pondérations q_j forme un nuage N(V) contenu dans $\mathbf{R}^n$. Contrairement au nuage $N(\Omega)$ dont le centre de gravité se situait à l'origine de $\mathbf{R}^p$ le centre de gravité du nuage N(V) n'est pas à l'origine de $\mathbf{R}^n$. On voit donc que la symétrie qui existait pour le tableau de contingence n'est plus vraie ici.

Les métriques

Pour mesurer la proximité entre individus, on munit l'espace des individus de la métrique quadratique définie par la matrice diagonale de terme général q_j où q_j correspond à l'importance donnée à la variable j. On a donc

$$d^2(x_i, x_{i'}) = \sum_{j \in V} q_j (x_i^j - x_{i'}^j)^2$$

Dans cette somme, il peut exister des valeurs de j pour lesquelles les variables correspondantes ont des échelles très diverses. Si on désire faire jouer à chaque variable le même rôle dans la définition des proximités entre individus, on peut alors définir les valeurs q_j de la manière suivante:

$$q_j = 1/\sigma_j^2,$$

σ_j étant l'écart-type de la variable j, c'est-à-dire,

$$\sigma_j^2 = \sum_{i \in \Omega} p_i (x_j^i)^2.$$

Cela revient à transformer le tableau initial en divisant chaque colonne j par σ_j et à prendre la métrique euclidienne usuelle.

Pour mesurer la proximité entre variables, on munit l'espace des variables de la métrique des poids D_p. On a

$$d^2(x^j,x^{j'}) = \sum_{i \in \Omega} p_i (x_i^j - x_i^{j'})^2.$$

Les variables étant centrées par hypothèse on obtient les relations suivantes liant la distance aux variances et covariances des variables j et j':

$$d^2(x^j,x^{j'}) = \mathrm{var}(x^j) + \mathrm{var}(x^{j'}) - 2\,\mathrm{cov}(x^j,x^{j'})$$

et

$$d^2(0,x^j) = \mathrm{var}(x^j).$$

Dans le cas où les données sont normalisées on obtient cette fois une relation liant la distance à la corrélation des variables j et j':

$$d^2(x^j,x^{j'}) = 2(1-\mathrm{cor}(j,j')) \text{ et } d^2(0,x^j) = 1$$

Toutes les variables sont sur la sphère de rayon 1 de l'espace $\mathbf{R}^n$. Remarquons que deux variables fortement corrélées seront très proches si la corrélation est positive et au contraire très éloignées si la corrélation est négative. Deux variables faiblement corrélées seront à distance moyenne.

Information apportée par le tableau

On a défini précédemment deux nuages $N(\Omega)$ et $N(V)$. Calculons l'inertie du premier par rapport à l'origine qui est aussi le centre de gravité:

$$\begin{aligned} I(\Omega,p_i,q_j) &= \sum_{i \in \Omega} p_i d^2(0,x_i) \\ &= \sum_{i \in \Omega} p_i \sum_{j \in V} q_j (x_i^j)^2 \\ &= \sum_{i \in \Omega} \sum_{j \in V} p_i q_j (x_i^j)^2 \end{aligned}$$

On peut de la même facon calculer l'inertie du nuage $N(V)$ par rapport à l'origine:

$$\begin{aligned} I(V,q_j,p_i) &= \sum_{j \in V} q_j d^2(0,x^j) \\ &= \sum_{i \in \Omega} \sum_{j \in V} q_j p_i (x_i^j)^2 \end{aligned}$$

Les deux inerties sont donc égales. Nous retiendrons cette quantité comme mesure de l'information apportée par le tableau. On associera à tout tableau $X(\Omega,V)$ muni des pondérations $(p_1, \ldots, p_n)$ et $(q_1, \ldots, q_p)$ cette information que l'on notera

$$I(X(\Omega,V),(p_i),(q_j)) = \sum_{i \in \Omega} \sum_{j \in V} p_i q_j (x_i^j)^2$$

Tableau associé à un couple de partitions

Ayant un tableau X initial défini sur les deux ensembles Ω et V, nous allons associer à une partition P de W et une partition Q de V un nouveau tableau. Soient $P = (P_1, \ldots, P_K)$ une partition de Ω en K classes et $Q = (Q_1, \ldots, Q_L)$ une partition de V en L classes.

On définit le tableau X(P,Q):
$\forall k = 1, \ldots, K$ et $\forall \ell = 1, \ldots, L$

$$x_k^\ell(P,Q) = \left(\sum_{i \in P_k} \sum_{j \in Q_\ell} p_i q_j x_i^j \right) \left(\sum_{i \in P_k} \sum_{j \in Q_\ell} p_i q_j \right)^{-1}$$

x_k^ℓ représente la moyenne des valeurs x_i^j correspondant au couple (P_k,Q_ℓ) pondérées par les poids $p_i q_j$.

On associe au nouveau tableau X(P,Q) les pondérations

$$p'_k = \sum_{i \in P_k} p_i, \quad \forall\, k=1, \ldots, K,$$

$$q'_\ell = \sum_{i \in Q_\ell} q_j, \quad \forall\, \ell=1, \ldots, L.$$

On peut justifier ces choix en montrant que le tableau X(P,Q) peut être obtenu grâce aux deux étapes suivantes:

Soient X(Ω,Q) le tableau associé au nuage des individus Ω cette fois mesurés par les L nouvelles variables obtenues en remplacant les p variables initiales par les L variables moyennes de chaque classe.

$$x_i^\ell(\Omega,Q) = \left(\sum_{j \in Q_\ell} q_j\, x_i^j \right) \left(\sum_{j \in Q_\ell} q_j \right)^{-1}$$

et

$$q'_\ell = \sum_{j \in Q_\ell} q_j$$

Maintenant, si dans ce nouveau nuage, on remplace les n individus par les K centres de gravité des classes de P, on retrouve alors la définition de x_k^ℓ

$$x_k^\ell(P,Q) = \left(\sum_{i \in P_k} p_i\, x_i^\ell(\Omega,Q) \right) \left(\sum_{i \in P_k} p_i \right)^{-1}$$

$$= \left(\sum_{i \in P_k} \sum_{j \in Q_\ell} p_i\, q_j\, x_i^j \right) \left(\sum_{i \in P_k} \sum_{j \in Q_\ell} p_i\, q_j \right)^{-1}$$

et $p_k = \sum_{i \in P_k} p_i$.

Si on avait effectué les mêmes opérations de manière symétrique en commencant par regrouper les classes de P puis celles de Q, on obtiendrait évidemment le même résultat.

Le regroupement des lignes et des colonnes conduit donc à une perte d'information. On cherchera à minimiser cette perte.

Relation entre inertie intraclasse et information

Considérons le nuage des individus obtenus après regroupement des variables selon la partition Q; il est défini par:

$$\{ z_i = (z_i^1, \ldots\ z_i^L)\ ,\ \forall i \in \Omega \}$$

où $z_i^\ell = \left(\sum_{i \in Q_\ell} q_j x_i^j \right) \left(\sum_{j \in Q_\ell} q_i \right)^{-1}$.

La pondération des individus est toujours (p_i).

La métrique est définie par la matrice diagonale de terme général q'_ℓ où

$$q'_\ell = \sum_{j \in Q_\ell} q_j$$

Notons alors W(P/Q) l'inertie intraclasse de la partition P du nuage que l'on vient de définir:

$$W(P/Q) = \sum_{k=1}^{K} \sum_{i \in P_k} p_i \sum_{\ell=1}^{L} q'_\ell\, (z_i^\ell - g_k^\ell)^2$$

où $g_k^\ell = \left(\sum_{i \in P_k} p_i z_i^\ell \right) \left(\sum_{i \in P_k} p_i \right)^{-1}$.

De la même façon, on note W(Q/P) l'inertie intraclasse de la partition Q du nuage des variables obtenu après regroupement des individus. On peut alors montrer en utilisant le théorème de Huygens que ces inerties intraclasses sont reliées simplement aux mesures d'information des tableaux X(P,Q), X(Ω,Q) et X (P,V).

Indice d'interprétation d'un couple de partitions

En utilisant les résultats du paragraphe dans lequel on considère le nuage N(IxJ) dans **R**, on définit les indicateurs suivants:

À chaque couple de classes (k,ℓ), on peut associer la somme des inerties des couples (i,j) appartenant au couple (k,ℓ), par rapport à l'origine.

$$T_{k\ell} = \sum_{i \in P_k} \sum_{j \in Q_\ell} p_i\, q_j\, (x_i^j)^2$$

On définit l'inertie conservée par le centre de gravité g_k^ℓ de la classe (k,ℓ):

$$B_{k\ell} = p_k\, q_\ell\, (g_k^\ell)^2$$

$W_{k\ell}$ est l'inertie de la classe (k,ℓ). On a

$$T_{k\ell} = B_{k\ell} + W_{k\ell}$$

$$I(I,J) = T = \sum_{k,\ell} T_{k\ell}$$

$$I(P,Q) = B = \sum_{k,\ell} B_{k\ell}$$

Ces relations permettent de définir:

- le tableau des valeurs $B_{k\ell}/B$ représentant la part d'inertie conservée par la classe sur l'inertie totale conservée.
- le tableau des valeurs $B_{k\ell}/T_{k\ell}$ représentant la part d'inertie conservée par la classe, par rapport à l'inertie initiale des points de la classe. Cette quantité indique la qualité de représentation d'une classe.

2.6.7 Tableaux binaires

On se place cette fois exactement dans les conditions du petit exemple décrit dans l'introduction. L'objectif est d'obtenir en réordonnant les lignes et les colonnes du tableau initial suivant les deux partitions, des blocs homogènes de 1 ou de 0. Pour ceci, nous associons à chaque couple de partitions, un tableau binaire "idéal" de dimension (K, L) où K et L sont les nombres de classes des partitions P et Q. L'algorithme proposé recherche en même temps un couple de partitions P et Q en K et L classes et un tableau binaire idéal optimisant un critère qui mesure l'écart entre le tableau initial structuré suivant les partitions et le tableau binaire idéal. En plus des deux partitions obtenues, ce tableau binaire constitue un résumé simple à analyser du tableau de données initial.

Pour justifier et préciser la recherche de partitions simultanées d'un tableau binaire, on peut reprendre les propositions faites par Lerman à propos de la notion de classe polythétique (LERMAN 1981). Il rappelle d'abord la notion de classe polythétique selon Becker: "une classe polythétique G d'une classification naturelle se réfère à un sous-ensemble B d'attributs tel que:

a) Chaque élément de la classe possède une proportion importante d'attributs de B.
b) Chaque attribut de B est présent dans une proportion importante d'éléments de G.
c) Il n'y a pas nécessairement un attribut qui soit possédé par tous les éléments de G".

Lerman généralise alors cette notion: "la situation plus générale d'une bonne classification sur E avec une bonne classification sur A conformément à la définition approximative de Becker est celle où à chacune des classes E_i de la partition $(E_1, ..., E_k)$ correspond une réunion B de classes de la partition $(A_1, ..., A_k)$ sur A, à laquelle la classe E_i se réfère. Réciproquement à toute classe A_j correspond une réunion G de classes E_i à laquelle la classe A_j se réfère".

La situation ainsi décrite peut être représentée par le tableau binaire réordonné suivant les partitions. L'objectif étant de trouver des blocs homogènes, c'est-à-dire remplis soit de 1, soit de 0, on associe à chaque couple (k,ℓ) de classes une valeur binaire idéale (1 ou 0). On obtient ainsi un tableau binaire que l'on appelle noyau. On cherche alors à minimiser le nombre de fois où la valeur associée à un couple (i,j) est différente de la valeur idéale associée au couple de classes auquel appartient (i,j). Cette quantité représente l'écart entre le tableau initial et le tableau idéal.

Exemple: Soit le tableau binaire

	J								
I	1	1	0	1	1	1	1	0	0
	0	0	1	0	0	1	1	0	0
	1	1	1	1	1	0	1	0	1
	0	0	0	1	1	0	1	1	1
	0	1	0	1	1	1	1	1	1

Soit le couple de partitions

$$\begin{cases} (\{1,2,3\}, \{4,5\}) \text{ pour I} \\ (\{1,2,3\}, \{4,5,6,7\}, \{8,9\}) \text{ pour J} \end{cases}$$

Soient les valeurs idéales suivantes

1	1	0
0	1	1

Alors l'écart entre le tableau initial et le tableau idéal est égal à 9.

On notera n et p les cardinaux des deux ensembles I et J. Le tableau binaire sera noté (x_i^j). On a donc $x_i^j \in \{0,1\}$. Dans toute la suite, les partitions P et Q seront des partitions des deux ensembles I et J en K et L classes au plus. K et L sont des constantes fixées arbitrairement. Soit $\mathbf{L}_{K,L}$ l'ensemble des tableaux binaires à K lignes et L colonnes.

$$L \in \mathbf{L}_{K,L} \Leftrightarrow L = (a_k^\ell)$$

où $a_k^\ell \in \{0,1\}$, $\forall$ k=1, K et $\forall$ ℓ=1, L.

Par abus de notation, L représente à la fois le nombre de classes demandées en colonne et un noyau. Le contexte permettra facilement d'identifier les deux cas.

Le critère

On a vu précédemment que le critère que l'on cherche à minimiser représente l'écart entre le tableau initial structuré suivant les 2 partitions P et Q et le tableau idéal L. Ce critère peut s'exprimer comme une fonction des deux partitions P et Q, et du noyau L:

$$W(P,Q,L) = \sum_k \sum_\ell \sum_{i \in P_k} \sum_{j \in Q_\ell} |x_i^j - a_k^\ell|$$

L'algorithme

Plusieurs algorithmes sont possibles pour optimiser le critère. Nous avons retenu l'algorithme suivant:

Partant d'un élément (P,Q,L) initial, on fixe Q et on cherche à améliorer P et L, puis on fixe P et on cherche à améliorer Q et L. On construit ainsi une suite (P^n,Q^n,L^n) qui fait décroître le critère W.

Résultat sur un exemple

Si on reprend le petit exemple du paragraphe 2.6.2, l'algorithme proposé fournit, lorsqu'on demande 3 classes en ligne et 2 classes en colonne les résultats suivants:

	1	3	5	8	10	2	4	6	7	9
c	1	0	0	1	0	0	0	0	0	1
g	0	0	0	1	1	1	0	0	0	0
i	1	0	0	0	1	0	1	0	0	0
a	1	1	1	1	1	0	0	0	0	0
d	1	1	0	1	0	0	0	0	0	0
h	1	1	1	1	1	0	0	1	0	1
b	0	0	0	0	0	1	1	1	1	1
e	0	0	0	0	0	1	1	1	1	1
f	0	0	0	0	0	1	0	1	1	1
j	0	0	0	0	0	1	1	0	1	0

Tableau initial réordonné

	1	2
1	0	0
2	1	0
3	0	1

Tableau des valeurs idéales

	1	2
1	6	3
2	13	2
3	0	17

Tableau P, Q

	1	2
1	60	80
2	87	87
3	100	85

Homogénéité par classe

On a bien retrouvé les résultats prévus et de plus, l'algorithme obtient ces résultats pour les 5 tirages. Le tableau des "valeurs idéales" permet de caractériser les classes entre elles, tandis que le tableau initial réordonné permet d'analyser plus finement les partitions. Enfin le dernier tableau permet de mesurer la qualité de représentation de chaque couple de classes. Par exemple, le couple (classe 3 en ligne et classe 1 en colonne) est parfaitement représenté par la valeur 0 (100%), alors que le couple classe 1 en ligne, classe 1 en colonne est très moyennement représenté par la valeur 0 puisque seulement 60% des valeurs sont 0.

2.7 Classification d'un tableau de proximités

2.7.1 Introduction

Le tableau de données de type Individus × Variables est le tableau de données le plus communément rencontré en analyse des données. Cependant dans certains cas l'utilisateur recueille

directement de son appareil de mesure un tableau de proximités sur un ensemble d'objets. Par exemple, le tableau de confusion entre les syllabes émises et celles reconnues par une autre personne est un tableau de proximité sur l'ensemble des syllabes. On peut aussi construire ce genre de tableau à partir d'un tableau de type Individus × Variables. La matrice des valeurs absolues des corrélations entre variables est un bon exemple car aucune méthode de classification basée sur le centre de gravité ne peut traiter ce tableau. On peut aussi définir une distance non euclidienne entre les individus et dans ce cas les méthodes de classification classiques sont inopérantes.

2.7.2 Méthode de classification d'un tableau de proximités

On suppose que le tableau de proximités entre les individus de Ω est un tableau de mesures de dissemblances (ou indices de distances). Ce tableau peut être aussi un tableau de distances entre ces individus. On notera d la mesure de dissemblance. La méthode présentée est de type Nuées Dynamiques.

L'algorithme des Nuées Dynamiques

Rappelons que d'une manière générale la méthode des Nuées Dynamiques permet d'optimiser le critère suivant:

$$W(P,L) = \sum_{\ell=1}^{k} D(P_\ell, L_\ell)$$

où $P = (P_1, ..., P_k)$ est une partition en k classes de Ω, $L = (L_1, ..., L_k)$ un élément de $\mathbf{L}_k$ et D une mesure d'adéquation.

On définit une masse μ sur Ω

$$\mu: \Omega \to \mathbf{R}^+.$$

Soit D la fonction de $P(\Omega) \times P(\Omega)$ dans $\mathbf{R}^+$ telle que

$$D(P_\ell, L_\ell) = \sum_{w \in P_\ell} \mu(w) \sum_{a \in L_\ell} d(w,a)$$

$\mathbf{L}$ étant inclus dans $P(\Omega)$, on utilise cette fonction comme mesure d'adéquation entre $P(\Omega)$ et $\mathbf{L}$ d'où l'écriture du critère

$$W(P,L) = \sum_{\ell=1}^{k} \sum_{w \in P_\ell} \mu(w) \sum_{a \in L_\ell} d(w,a).$$

Chaque élément w de Ω est affecté à la classe P_ℓ si $D(\{w\}, L_\ell)$ est minimum. En cas d'égalité on choisit la classe de plus petit indice. La nouvelle partition $P = (P_1, ..., P_k)$ est donc définie par:

$$P_\ell = \{w \in \Omega:\ D(\{w\}, L_\ell) \leq D(\{w\}, L_{\ell'}) \quad \text{pour} \quad \ell \neq \ell'\}$$

À chaque classe P_ℓ est associée une nouvelle représentation L_ℓ telle que

$$D(P_\ell, L_\ell) = \min_{L \in \mathbf{L}} D(P_\ell, L)$$

Ceci revient à sélectionner les m éléments a de Ω minimisant $D(P_\ell, \{a\})$ puisque

$$D(P_\ell, L) = \sum_{a \in L} \sum_{w \in P_\ell} \mu(w) d(w,a) = \sum_{a \in L} D(P_\ell, \{a\})$$

En cas d'égalité on choisit l'élément de plus petit indice. L'algorithme des Nuées Dynamiques ainsi défini fait décroître le critère à chaque itération et la suite $v_n = (P_n, L_n)$ converge dans $\mathbf{P}_k \times \mathbf{L}^k$ en un nombre fini d'itérations.

Initialisation

L'initialisation peut se faire, comme pour tout algorithme des Nuées Dynamiques, en choisissant (au hasard ou non) la partition initiale ou les éléments de la représentation initiale. Cependant, on propose ici une procédure d'initialisation qui permet d'obtenir des solutions en 2, 3, ..., k classes. Au départ on sélectionne l'individu w_1 tel que

$$D(\{w_1\}, \Omega) = \min_{w \in \Omega} D(\{w\}, \Omega)$$

puis l'individu w_2 tel que

$$d(w_1, w_2) = \max_{w \in \Omega} d(w_1, w)$$

À chacun des individus w_1 et w_2 on associe les (m-1) individus les plus proches. On obtient ainsi deux représentations L_1 et L_2. Partant de

L_1 et L_2, à la convergence de l'algorithme décrit ci-dessus, nous avons une partition en deux classes. On sélectionne alors l'individu w_3 le plus mal classé au sens de D; c'est-à-dire l'individu w_3 tel que:

$$D(\{w_3\}, L_{(w_3)}) = \underset{w \in \Omega}{\text{Max}}\ D(\{w\}, L_{(w)})$$

où $L_{(w)}$ est le représentant de la classe contenant w.

Puis on recherche les (m-1) éléments les plus proches de w_3 pour construire L_3. On réinitialise l'algorithme des Nuées Dynamiques en ajoutant aux représentants obtenus à la convergence précédente ce nouveau représentant. On recommence la procédure jusqu'à la solution en k classes. Cette stratégie permet d'avoir une solution pour 2, 3, ..., k classes mais ne donne pas le moyen de faire un choix entre ces solutions.

2.7.3 Méthode optimisant un critère indépendant du nombre de classes

Position du problème

Une des solutions permettant de résoudre le problème du choix du nombre de classes consiste à ajouter au critère à minimiser, le coût de création d'une classe. En imposant ce coût, nous devons résoudre un problème similaire au problème de la localisation des centres (COOPER 1972), (DUNSTAN, WELSH 1973). Sous sa forme la plus simple, ce problème de localisation des centres peut s'énoncer de la façon suivante:

Soit un ensemble Ω de n clients répartis dans un territoire et devant être desservis par un ensemble de k centres $L = \{y_1, ..., y_k\}$ on cherche à déterminer le nombre optimal de centres, et les centres choisis. À chaque centre y_j on associe le coût c_j, qui est le coût, d'implantation de ce centre. À chaque client x_i on associe le coût $d(w_i, y_j)$ qui est le coût d'affectation du client w_i au centre y_j. Le coût d'une solution N (sous-ensemble de Ω) est la somme de deux coûts:

- le coût des centres:

$$w_2(N) = \sum_{y_j \in N} c_j$$

- le coût d'affectation des clients aux centres $y_j \in N$:

$$W_1(N) = \sum_{w_i \in w} \underset{y_j \in N}{\text{Min}} \{d(w_i, y_j)\}$$

Le problème est de trouver un sous-ensemble N^* de Ω tel que:

$$W(N^*) = W_1(N^*) + W_2(N^*)$$

$$= \underset{N \in L}{\text{Min}} \{W_1(N) + W_2(N)\}$$

La solution optimale de ce problème est le meilleur compromis possible entre les deux solutions extrêmes:

- La solution N=L où le coût d'affectation est minimum, mais le coût de création des centres est maximum. Avec cette solution tous les clients sont choisis comme centres.
- La solution $N=\{y_j\}$ où y_j est le centre de coût minimum qui réduit au maximum le coût de construction des centres mais augmente le coût d'affectation des clients.

La recherche de la solution optimale par l'énumération complète des solutions possibles est difficilement réalisable avec un ensemble de clients supérieur à 100 car le nombre de solutions possibles est de l'ordre de n!. Ainsi la recherche de solutions sous-optimales s'avère nécessaire. Dans les applications, on utilise fréquemment les algorithmes gloutons (DUSTAN, WELSH 1973 , MINOUX 1977). Ces méthodes réduisent ce problème à un problème de cheminement dans le graphe associé au treillis des parties de l'ensemble Ω des clients, mais la complexité de ces algorithmes est de l'ordre de n^3. Une autre approche peut être développée en supposant que l'ensemble des centres est inclus dans l'ensemble des clients. Plus précisément l'ensemble des sites d'implantation est inclus dans les sites où les clients exercent leurs activités. Cette contrainte n'est pas, en pratique, bien restrictive d'autant que l'on peut créer un site client fictif à chaque site d'implantation possible d'un centre.

Cette approche permet d'utiliser l'algorithme des Nuées Dynamiques.

2.8 Interprétation d'une partition

Dans ce paragraphe, nous présentons des indices de description d'une partition par un ensemble de variables. Ces indices sont fondés pour la plupart sur la décomposition de l'inertie de la population en inertie interclasse et inertie intraclasse associées à la partition étudiée. Les variables de description peuvent être intervenues dans la détermination de la partition. On parle alors de variables actives. Elles peuvent aussi ne pas être intervenues dans la détermination de la partition. On parle alors de variables passives ou illustratives. Cela étant, les indices proposées sont particulièrement adaptés à l'analyse de partitions obtenues par des algorithmes de partitionnement optimisant un critère d'inertie.

2.8.1 Notations et définitions

Soit $W = (w_1, ..., w_n)$ l'ensemble des individus de la partition à décrire $P = (P_1, ..., P_k)$. Si les p variables de description sont quantitatives, on se place dans le nuage $N(\Omega)$ de $\mathbb{R}^p$ associé. Si les variables sont qualitatives, on se place dans le nuage $N(\Omega)$ de $\mathbb{R}^p$ où p est le nombre total de modalités. On obtient ce nuage en codant les variables sous forme disjonctive complète. Les indices proposés reposent sur la décomposition de l'inertie totale en inertie interclasse et inertie intraclasse. Comme on l'a vu au paragraphe 1.3.4, cette décomposition est vraie globalement, par variable, par classe et par couple (variable, classe).

$$T = B+W$$

$$T^j = B^j+W^j$$

$$T_l = B_l+W_l$$

$$T^j_l = B^j_l+W^j_l$$

Rappelons que la métrique M utilisée pour le calcul des inerties est diagonale.

2.8.2 Indice général

La part d'inertie conservée ou expliquée en assimilant les individus aux k centres de gravité s'écrit:

$$R = \frac{B}{T}$$

Plus R est fort, plus la partition constitue une bonne représentation des individus. Ainsi, si R = 1 cela signifie que le nuage des individus se réduit aux nuages des k centres de gravité des classes.

Remarquons que, lorsque la partition est obtenue par l'algorithme des centres mobiles sur les variables de description, le critère minimisé est W. Cet algorithme cherche donc à maximiser R.

2.8.3 Contributions des variables

Pour chaque variable j, on définit un indice analogue à R:

$$COR(j) = \frac{B^j}{T^j}.$$

Il représente la part d'inertie de la variable j pris en compte par la partition. Cet indice mesure le pouvoir discriminant de la variable j par rapport à la partition. Il est intéressant de comparer les valeurs de COR(j) avec R; en effet si COR(j) > (resp <) R, la variable j est plus (resp. moins) discriminante que la moyenne puisque R représente le pouvoir discriminant "moyen" des variables vis-à-vis de la partition P.

L'indice suivant est un indice complémentaire à COR(j) et doit être interprété en regard de ce dernier.

Pour j = 1, ..., p il s'agit de la contribution relative de la variable j à l'inertie interclasse de la partition:

$$CTR(j) = \frac{B^j}{B}.$$

On a

$$\sum_{j=1}^{p} CTR(j) = 1.$$

Ces deux indices ont tendance à varier dans le même sens. C'est vrai en particulier dans le cas, courant avec des critères d'inertie, où la partition a des classes de volumes analogues. Le cas plus rare où l'on a COR(j) faible (resp. fort) et CTR(j) fort (resp. faible) est intéressant: cela signifie que la variable j malgré une forte (resp. faible) contribution à l'inertie de P est peu (resp. très) discriminante.

2.8.4 Description des classes

Pour chaque classe ℓ, on définit:

$$T(\ell) = \frac{T_\ell}{T}$$

qui représente le pourcentage d'inertie extrait par la classe ℓ;

$$B(\ell) = \frac{B_\ell}{B}$$

qui représente la contribution relative de la classe ℓ, à l'inertie interclasse et

$$W(\ell) = \frac{W_\ell}{W}$$

qui représente la contribution relative de la classe ℓ, a l'inertie intraclasse. On a

$$\sum_{\ell=1}^{k} T(\ell) = \sum_{\ell=1}^{k} B(\ell) = \sum_{\ell=1}^{k} W(\ell) = 1$$

$B(\ell)$ indique la position de la classe ℓ par rapport au centre de gravité global. Plus cet indice est fort, plus la classe ℓ occupe une place excentrée. $W(\ell)$ indique la concentration de la classe ℓ indépendemment de sa position. Il doit être examiné en regard du cardinal de la classe ℓ. Une valeur de $W(\ell)$ faible sera d'autant plus significatif de la concentration de la classe ℓ que son cardinal sera élevé.

2.8.5 Description des classes par variable.

Pour chaque variable j et chaque classe ℓ, on définit:

$$COR(j,\ell) = \frac{B^j_\ell}{T^j}$$

qui représente le pourcentage du pouvoir discriminant de la variable j pris en compte par la classe ℓ; en effet on a:

$$\sum_{\ell=1}^{k} COR(j,\ell) = COR(j)$$

$COR(j,\ell)$ mesure l'adéquation entre la classe ℓ et la variable j. Plus $COR(j,\ell)$ est fort, plus la variable j aura un comportement homogène sur les éléments de la classe ℓ. L'examen de cet indice est particulièrement intéressant pour les variables à fort pouvoir discriminant: il indique la répartition entre les classes de ce pouvoir discriminant.

2.8.6 Indices complémentaires

Pour chaque variables j et chaque classe ℓ, on définit:

$$DIS(j,\ell) = (g^j_\ell - g^j)/\sigma_j,$$

g étant le centre de gravité du nuage, g_ℓ étant le centre de gravité de la classe P_ℓ, σ_j étant l'écart-type de la variable j. Cela permet de voir rapidement les variations de la variable j en fonction des classes et réciproquement le comportement de la classe ℓ par rapport aux variables.

On définit également l'indice

$$CE(j,\ell) = \frac{B^j_\ell}{B}$$

qui représente la contribution relative de la classe ℓ et de la variable j à l'inertie interclasse. Plus cet indice est fort, plus il indique que pour la variable j la classe ℓ occupe une position excentrée.

On définit:

$$CTR(j,\ell) = \frac{B^j_\ell}{B_\ell}$$

qui représente la contribution relative de la classe ℓ à l'inertie interclasse de la variable j. On a

$$\sum_{\ell=1}^{k} CTR(j,\ell) = 1.$$

Plus cet indice est fort, plus il indique que la classe ℓ joue un rôle important dans l'inertie extraite par la partition pour la variable j. Ces in-

dices ne sont utiles que si l'on désire étudier de manière détaillée une classe particulière ou une variable particulière.

2.8.7 Pour le cas particulier des variables qualitatives

Indiquons que tous les indices décrits précédemment sont à lire en remplaçant le terme "variable" par "modalité d'une variable". À partir des contributions relatives de chaque modalité j d'une variable q, on définit la contribution relative de la variable q:

$$CTR(q) = \sum_{j \in q} CTR(j) = \sum_{j \in q} \frac{B^j}{B}$$

Pour étudier la liaison entre une classe ℓ et une modalité j, on considère le khi-deux du tableau de contingence croisant les deux variables qualitatives à deux modalités indicatrices de ℓ et de j. On le note CHI(j,ℓ).

2.8.8 Exemple d'interprétation d'une partition

Variables quantitatives

Cet exemple a trait à une classification en 4 classes des données de Ruspini (cf. DIDAY 1980). Il s'agit de 75 points du plan bien structurés en 4 classes.

$$R = 0.80$$

Commentaire: la partition extrait 80% de l'inertie totale. Elle est donc bien représentative des données de départ.

Coordonnées des centres de gravité des classes

Variable	Population	Classe 1	Classe 2	Classe 3	Classe 4
(effectif)	(75)	(23)	(5)	(10)	(37)
var1	39.6	42.3	84.2	85.7	17.4
var2	49.0	30.7	87.2	13.1	64.9

Écart-types des classes pour chaque variable

Variable	Population	Classe 1	Classe 2	Classe 3	Classe 4
(effectif)	(75)	(23)	(5)	(10)	(37)
var1	28.5	18.1	8.0	12.4	7.4
var2	25.2	13.3	6.9	9.3	11.8

On voit par exemple que la variable 1 ne permet pas de distinguer les classes 2 et 3 et que par contre la variable 2 distingue bien les 4 classes entre elles. La classe 4 est bien concentrée pour la variable 1 et que la classe 2 est bien concentrée pour la variable 2.

Rôle des variables dans la partition:

COR(J) = B(J,.)/T(J,.) → Pourcentage du pouvoir discriminant de la variable J.

CTR(J) = B(J,.)/B(.,.) → Contribution relative de la varible J à l'inertie interclasses.

Variable	COR	CTR
var1	81.2	57.1
var2	78.2	47.9

Ici COR(J) est proche de R (80%): les variables ont un pouvoir discriminant comparable. CTR(J) n'apporte pas d'éléments nouveaux par rapport a COR(J).

Indices de description des classes de la partition:

T(K) = T(.,K)/T(.,.) → Pourcentage d'inertie extraite par la classe K.

B(K) = B(.,K)/B(.,.) → Contribution relative de la varible K à l'inertie interclasses.

W(K) = W(.,K)/W(.,.) → Contribution relative de la varible K à l'inertie intraclasses.

Classe	Effectif	T(K)	B(K)	W(K)
1	23	18.1	9.2	53.3
2	5	16.8	20.4	2.50
3	10	34.5	40.5	11.0
4	37	30.6	29.9	33.1

La faible valeur de l'inertie de la classe 2 (W(2)) s'explique surtout par son faible effectif. La classe 3 est la plus éloignée du comportement moyen des points (B(3) fort).

	Partition		Classe 1		Classe 2		Classe 3		Classe 4	
Var	COR	CTR	COR	CTR	COR	CTR	COR	CTR	COR	CTR
Var1	81.2	57.1	0.5	3.9	17.1	58.8	36.4	63.3	27.2	64.0
Var2	78.2	42.9	16.1	96.1	15.3	41.2	27.3	36.7	19.7	36.0

Indices décrivant les rôles d'une variable et d'une classe:

COR(J) = B(J,.)/T(J,.) → Pourcentage du pouvoir discriminant de la variable J.

COR(J;K) = B(J,K)/T(J,.) → Pourcentage du pouvoir discriminant de la variable J pris en compte par la classe K.

CTR(J) = B(J,.)/B(.,.) → Contribution relative de la variable J à l'inertie interclasses.

CTR(J,K) = B(J,K)/B(.,K) → Contribution relative de la variable J et la classe K à l'inertie inter-classes.

Sur ce tableau synthétique, on voit par exemple le rôle important de la variable 1 dans la formation des classes 3 et 4 et le rôle important de la variable 2 dans la formation de la classe 1.

Tableau des indices de dispersion des variables des classes par rapport au nuage (=DD):

DD → (CGP(J,K)-(G(J))/SIGMA(J)

Avec:

CGP(J,K) → Centre de gravité de la classe K pour la var J.

CG(J) → Centre de gravité du nuage pour la var J.

SIGMA(J) → Écart-type du nuage pour la var J.

Variable	Classe 1	Classe 2	Classe 3	Classe 4
var1	0.129	1.529	1.652	-0.743
var2	-0.725	1.517	-1.426	0.631

Représentation graphique des indices de dispersion DIS des variables pour chaque classe:

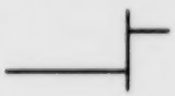
De plus les indices CE et CTR (j,k) sont édités en colonne:

CE = B(J,K)/B(.,.) → Contribution relative de la classe et de la variable à l'inertie inter-classes.

CTR-K = B(J,K)/B(.,K) → Contribution relative de la variable à l'inertie inter de la classe.

Classe 1

	CE	CTR-K	-1.5 -1.2 -0.9 -0.6 -0.3 0.0 0.3 0.6 1.2 1.5 1.9
Var1	0.6	3.9	
Var2	8.5	96.1	



La classe 1 se distingue par des valeurs faibles pour la variable 2.

Représentation graphique des indices de dispersion DD des classes pour chaque variable. De plus les indices CE et CTR(J) sont édités en colonne ainsi que le cardinal de chaque classe (N0):

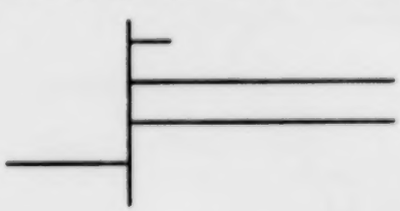
NC = Effectif de la classe

CE = B(J,K)/B(.,.) → Contribution relative de la classe et de la variable à l'inertie inter-classes.

CTR-J = B(J,K)/(B(J,.) → Contribution relative de la classe à l'inertie interclasses de la variable J.

Variable: var1

K	NC	CE	CTR-J	-1.5 -1.2 -0.9 -0.6 -0.3 0.0 0.3 0.6 1.2 1.5 1.9
1	23	0.4	0.6	
2	5	12.0	21.0	
3	10	25.6	44.8	
4	37	19.2	33.5	



La variable 1 oppose les classes 2 et 3 à la classe 4: elle prend des valeurs fortes sur les classes 2 et 3 et faibles sur la classe 4.

Variables qualitatives

Sélection pour chaque classe des modalités significatives au sens du khi-deux et édition pour chacune d'elles:

. La valeur du khi-deux entre la classe et la modalité.

. La probabilité associée (supérieure à 5%).

. L'effectif de la modalité dans la classe.

. La fréquence de la modalité dans la population totale.

. La fréquence de la modalité relative à

l'ensemble des individus ayant pris la modalité.

Pour chaque modalité significative, on vérifiera que l'effectif de la modalité n'est pas trop faible dans la classe. Si une modalité est significative pour une classe, sa fréquence dans la classe est supérieure à sa fréquence dans la population totale. On examinera avec soin l'écart entres ces deux fréquences.

2.9 Analyse d'une multi-partition

Un algorithme de type Nuées Dynamiques comme la plupart des algorithmes de partitionnement conduit à des optimums locaux du critère et la solution obtenue peut différer suivant la partition ou la représentation choisie à l'initialisation de l'algorithme. À l'issue de m passages d'un algorithme de partitionnement sur un même ensemble d'objets, on obtient donc m partitions $(P^1, ..., P^m)$ généralement non identiques. On appelle multi-partition un tel ensemble de m partitions ($m > 1$).

Représentation d'une multi-partition

À toute multi-partition $(P^1,...,P^m)$ est associé un tableau M à n lignes et m colonnes dont la case (i,j), intersection de la $i^{ème}$ ligne et de la $j^{ème}$ colonne, contient le numéro de la classe de la partition P^j auquel appartient l'objet i de W. À partir du tableau M, on construit aisément la matrice carrée A de dimension n, telle que: $\forall(i,i') \in \Omega \times \Omega$, $A_{ii'}$ est le nombre de fois où les objets i et i' ont été classés ensemble dans les partitions $(P^1, ..., P^m)$. Dans ce paragraphe, nous présentons différentes techniques pour tirer parti de l'information apportée par la multi-partition sur la structure en classes de Ω et dans certains cas s'affranchir du problème de dépendance des solutions vis-à-vis de l'initialisation d'un algorithme de partitionnement.

2.9.1 Les formes fortes

On désigne par formes fortes de la multi-partition $(P^1, ..., P^m)$, les classes de la partition intersection des m partitions $P^1, ..., P^m$. (cf. DIDAY 1972, 1980). Autrement dit, les formes fortes sont les classes d'objets qui ont toujours été classés ensemble dans chacune des m partitions P^1, ..., P^m. Les formes fortes sont avant tout d'une grande utilité pour extraire d'une population les regroupements d'objets les plus "significatifs". Dans cette optique, les formes fortes les plus intéressantes sont celles de cardinal relativement élevé. Il est également utile d'examiner comment les formes fortes se situent les unes par rapport aux autres.

Si q ($q << n$) est le nombre des formes fortes, de la même manière qu'on a considéré le tableau associé à une multi-partition, on peut considérer le tableau des formes fortes N à q lignes et m colonnes dont la case (i,j) intersection de la $i^{ème}$ ligne et de la $j^{ème}$ colonne contient le numéro de classe de la partition P^j auquel appartient un représentant de la forme forte i. De manière analogue à la définition de la matrice A, on définit la matrice carrée B de dimension q: $\forall(f, f') \in F \times F$, $B_{ff'}$ est le nombre de fois où les représentants de f et f' ont été classés ensemble dans les partitions $P^1, ..., P^m$.

2.9.2 L'arbre de longueur minimum sur les formes fortes

On verra qu'il y a équivalence entre l'ultramétrique sous-dominante et l'arbre de longueur minimum associé. Pratiquement, l'édition de l'arbre de longueur minimum des formes fortes pour la distance de la différence symétrique semble apporter une vision plus nette du tableau des formes fortes que l'arbre hiérarchique. En particulier lorsque la structure en classes de l'ensemble Ω est assez forte, ces classes sont associées à des directions de l'arbre de longueur minimum (cf. LECHEVALLIER 1974).

2.9.3 Partition centrale associée à une multi-partition

À partir d'une multi-partition (P^1, ..., P^m) obtenue après m passages d'un algorithme des Nuées Dynamiques, il peut être fructueux de construire une partition qui puisse s'interpréter comme un "consensus" entre les différentes partitions P^1, ..., P^m (cf. CELEUX 1984). La partition des formes fortes qui correspond à un consensus unanime est en général trop fine pour jouer ce rôle de partition consensus.

Considérations pratiques

• Lorsque l'ensemble des objets à classer possède une structure nette en k classes, la partition centrale obtenue la retrouve même si le nombre de classes demandé par l'utilisateur a été mal estimé (dans des proportions raisonnables).

• Lorsque la structure en classes est plus faible, mais réelle, la partition centrale trouvée sera un bon résumé des différents tirages et pourra être interprétée de manière privilégiée pour peu que le nombre de classes de départ soit pertinent.

• Lorsque la structure en classes est arbitraire, la partition centrale ne sera pas interprétable car possédant de trop nombreuses classes. Elle indiquera du moins le caractère ténu de toute classification.

2.10 La classification hiérarchique

2.10.1 Introduction

Les hiérarchies, par la commodité de leur interprétation visuelle constituent depuis longtemps une forme de classification très populaire. Les classifications "naturelles" des animaux et des végétaux sont des hiérarchies. Fréquemment, l'utilisateur est surtout intéressé par la détection de classes "bien significatives", issues de la hiérarchie, l'idéal étant que ces classes forment une partition obtenue par découpage de la hiérarchie selon une ligne horizontale bien placée.

Exemple: Le découpage (selon la ligne indiquée) de la hiérarchie définit la partition $P = \{P_1, P_2, P_3\}$ avec $P_1 = \{a, c\}$, $P_2 = \{b, f, d\}$, $P_3 = \{e, g\}$.

Rappelons qu'une hiérarchie H est un ensemble de parties appelées "paliers" de l'ensemble des individus Ω qui contient Ω et les singletons et dont l'intersection de deux parties est soit vide, soit identique à l'une d'entre elles. Il résulte de cette définition que dans une hiérarchie, chaque palier (non réduit à un singleton) est la réunion d'autres paliers. Par la suite nous utiliserons souvent la notion de hiérarchie binaire; on appelle ainsi une hiérarchie dont chaque palier est la réunion de 2 paliers.

Afin de pouvoir visualiser une hiérarchie par un graphique il faut "valuer" les paliers de la hiérarchie; autrement dit, leur associer une hauteur, d'où la notion de hiérarchie indicée que nous introduisons d'abord. Nous montrons ensuite que cette hiérarchie indicée est équivalente à une distance ultramétrique. Nous examinons plusieurs types d'indices d'agrégation utiles pour la construction d'une hiérarchie indicée. Nous présentons plusieurs algorithmes permettant de construire une hiérarchie indicée. On introduit enfin la notion d'arbre de longueur minimum et on montre son équivalence avec l'ultramétrique induite par la hiérarchie dite du "saut minimum".

2.10.2 Définition d'une hiérarchie indicée

Une hiérarchie indicée est un couple (H,f) où H est une hiérarchie et f une application de H dans $\mathbf{R}^+$ telle que:

1) f(h) = 0 si et seulement si h ne contient qu'un seul élément.

2) $\forall$ h et h' $\in$ H, $h \subset h' \wedge h \neq h' \Rightarrow f(h) < f(h')$.

Cet indice permet de tracer l'arbre hiérarchique associé et de définir un nouvel indice de dissimilarité sur Ω de la manière suivante:

$$\delta(k,\ell) = \text{Min } \{f(h):\ k,\ell \in h,\ h \in H\}$$

Plus les individus se regroupent bas dans l'arbre, plus ils se ressemblent au sens de cet indice.

2.10.3 Indice d'agrégation entre groupes d'individus

La construction d'une hiérarchie nécessite la connaissance d'une "mesure de ressemblance" entre groupes (cf. paragraphe 1.5) cette mesure est appelée "indice d'agrégation". C'est une application symétrique de $\mathbf{P}(\Omega) \times \mathbf{P}(\Omega)$ dans $\mathbf{R}^+$.

Considérons une hiérarchie binaire H et un indice d'agrégation δ. Soit f la fonction $f(h_i \cup h_j) = \delta(h_i, h_j)$ où $h_i, h_j \in H$. Pour les indices d'agrégation courants, (H, f) est une hiérarchie indicée. Lorsque ce n'est pas le cas (existence d'"inversions"), on peut utiliser la fonction $f(h_i \cup h_j) = \text{Max } (\delta(h_i, h_j), f(h_i), f(h_j))$ qui généralise la première et qui garantit que (H, f) est une hiérarchie indicée.

Citons quelques indices d'agrégation parmi les plus classiques.

L'indice d'agrégation du lien maximum

$$\delta_1(h_1, h_2) = \max_{\substack{w_i \in h_1 \\ w_j \in h_2}} d(w_i, w_j)$$

L'indice d'agrégation du lien minimum

$$\delta_2(h_1, h_2) = \min_{\substack{w_i \in h_1 \\ w_j \in h_2}} d(w_i, w_j)$$

Dans le cas euclidien, on utilise souvent l'indice suivant:

Indice de l'augmentation d'inertie

$$\delta_3(h_1, h_2) = I(h_1 \cup h_2) - I(h_1) - I(h_2)$$

où I(h) représente l'inertie de h.

On a (cf. paragraphe 1.5.5):

$$\delta_3(h_1, h_2) = \frac{p(h_1)\, p(h_2)}{p(h_1)+p(h_2)}\, d^2(G(h_1), G(h_2)).$$

où G(h) est le centre de gravité de h et où p(h) est le poids de h.

Remarquons que si l'on prend simplement la distance entre les deux centres de gravité (sans la pondération):

$$\delta_4(h_1,h_2) = d^2(G(h_1), G(h_2))$$

et

$$f(h_1,h_2) = \delta_4(h_1 \cup h_2),$$

(H,f) n'est pas une hiérarchie indicée.

L'indice de l'augmentation d'inertie est l'un des plus utilisés car, à chaque étape, on optimise le critère d'inertie intraclasse. Cependant, une partition obtenue en coupant la hiérarchie à un niveau donné, sera en général moins bonne au sens du critère d'inertie intraclasse que celles fournies par l'algorithme des centres mobiles.

2.10.4 Construction de hiérarchies indicées

Ayant choisi un indice d'agrégation entre groupes d'individus, on peut imaginer de nombreux algorithmes pour construire une hiérarchie sur Ω. Les algorithmes de classification descendante hiérarchique consistent à découper de manière itérative la population Ω en partition de plus en plus fines jusqu'à la partition des singletons. On peut par exemple partir de la meilleure partition à 2 classes de Ω: $P^o = (P_1^o, P_2^o)$ au sens d'un critère W donné dépendant de δ; par exemple si $W(P) = \delta(P_1,P_2)$, il faut trouver P^o qui maximise

W parmi les partitions à 2 classes de Ω; on recommence ensuite le procédé sur chacune des deux classes ainsi obtenues jusqu'à ce que toutes les classes soient réduites à des singletons. Nous n'insisterons pas plus dans cet ouvrage sur les algorithmes de classification descendante hiérarchique car ils sont beaucoup moins utilisés que les algorithmes ascendants que nous allons décrire maintenant.

L'algorithme général de la classification ascendante hiérarchique.

Cet algorithme (dit de la CAH) consiste à construire, à l'aide de l'indice d'agrégation δ choisi, une suite de partitions de moins en moins fines dont les classes forment la hiérarchie H cherchée. Il s'énonce de la façon suivante:

1. Partir de la partition P^o dont les classes sont réduites à un élément.
2. Construire une nouvelle partition en réunissant les deux classes de la partition précédente qui minimisent δ.
3. Recommencer le procédé en 2 jusqu'à ce que toutes les classes soient réunies en une seule.

Si à l'étape 2 il y a plus d'un couple de classes qui minimise δ, on en choisi un au hasard; il n'y a donc pas toujours unicité de la hiérarchie obtenue. On peut remarquer aussi que la hiérarchie donnée par cet algorithme est nécessairement binaire.

Pour visualiser cette hiérarchie, il est nécessaire de l'indicer; certains indices peuvent faire apparaître des inversions mais qu'il est toujours possible de construire à partir d'une hiérarchie indicée au sens "large", dans ce cas il est inutile de conserver h dans H c'est pourquoi on est conduit à faire une épuration de tous ces paliers. On obtient ainsi une hiérarchie indicée mais qui n'est plus nécessairement binaire.

Algorithme des voisins réciproques

La méthode des voisins réciproques permet d'accélérer considérablement l'exécution de l'algorithme de classification ascendante hiérarchique et permet même pour certains indices d'agrégation comme l'indice d'augmentation d'inertie de ne stocker que la matrice des données (DE RHAN 1980). Cet algorithme revient à agréger tous les voisins réciproques à chaque étape de l'algorithme classique de la CAH au lieu de prendre seulement le couple qui minimise δ. Deux paliers sont voisins réciproques s'ils sont mutuellement plus proches voisins.

Exemple: On considère un ensemble de 5 points dans un plan et le tableau des distances euclidiennes associées:

	w_1	w_2	w_3	w_4	w_5
w_1	0	1	2.8	4.2	3.8
w_2		0	2.6	3.8	2.9
w_3			0	1.9	2.2
w_4				0	2.3
w_5					0

Par l'algorithme de la CAH on regroupe à chaque étape les deux paliers les plus proches. Avec l'indice d'agrégation du lien minimum on obtient à chaque étape les tableaux de distances suivants:

	(W_1,W_2)	W_3	W_4	W_5
(W_1,W_2)	0	2.6	3.8	2.9
W_3		0	1.9	2.2
W_4			0	2.3
W_5				0

	(W_1,W_2)	(W_3,W_4)	W_5
(W_1,W_2)	0	2.6	2.9
(W_3,W_4)		0	2.2
W_5			0

	$\{W_1,W_2\}$	$\{W_3,W_4,W_5\}$
$\{W_1,W_2\}$	0	2.6
$\{W_3,W_4,W_5\}$		0

Par l'algorithme des voisins réciproques on regroupe en une seule étape w_1 et w_2 ainsi que w_3 et w_4 car ils sont voisins réciproques. On a ainsi gagner une étape.

2.10.5 La formule de récurrence de Lance et Williams

À chaque étape, la CAH nécessite le tableau des indices d'agrégation entre les classes alors formées. Pour les indices d'agrégation rencontrés, on montre que l'on peut mettre à jour ce tableau directement sans revenir au tableau des distances initiales en utilisant une formule de récurrence. Par exemple, dans le cas du lien minimum, on a

$$\delta(A\cup B, C) = \min [\delta(A,C), \delta(B,C)].$$

LANCE et WILLIAMS (1967) ont généralisé cette formule de récurrence. Elle s'écrit:

$$\delta(A, B\cup C) = a_1\delta(A,B) + a_2\delta(A,C) + a_3\delta(B,C) + a_4|\delta(A,B) - \delta(A,C)|$$

où $A, B, C \in \mathbf{P}(\Omega)$.

Les valeurs des coefficients pour les principaux indices d'agrégation sont:
lien minimum $a_1 = a_2 = 1/2$, $a_3 = 0$, $a_4 = -1/2$
lien maximum $a_1 = a_2 = 1/2$, $a_3 = 0$, $a_4 = 1/2$
augmentation d'inertie

$$a_1 = \frac{\text{card}(A) + \text{card}(B)}{T}$$

$$a_2 = \frac{\text{card}(A) + \text{card}(C)}{T}$$

$$a_3 = -\frac{\text{card}(A)}{T}$$

$$a_4 = 0$$

avec $T = \text{card}(A) + \text{card}(B) + \text{card}(C)$

2.10.6 Hiérarchies indicées et ultramétriques

Ce paragraphe est essentiellement consacré à un résultat important en classification hiérarchique: l'équivalence entre hiérarchie indicée et ultramétrique.

Définition et propriétés d'une ultramétrique

Rappelons (paragraphe 1.5.1) qu'une distance δ est dite ultramétrique si l'inégalité triangulaire est remplacée par l'inégalité ultramétrique:

$$\forall\ w,\ w',\ w'' \in \Omega,$$

$$\delta(w,w'') \le \text{Max}(\delta(w,w'), \delta(w',w'')).$$

On montre facilement qu'une condition nécessaire et suffisante pour qu'une distance soit ultramétrique est que tous les triangles soient isocèles avec la base plus petite que les côtés égaux.

Recherche d'une hiérarchie indicée en terme d'optimisation

Sachant qu'il y a équivalence entre une hiérarchie indicée et une ultramétrique, il est naturel de se poser le problème du choix de la meilleure hiérarchie sous la forme: "Chercher $\delta \in U$ optimisant $\Delta(d,\delta)$" où d est la dissimilarité choisie, U est l'ensemble des ultramétriques et Δ est une mesure de l'adéquation entre d et δ; différentes approches ont été proposées dans cette direction. Les trois principales sont les suivantes: recherche de l'ultramétrique sous-dominante, recherche de l'ultramétrique sur-dominante, recherche de l'ultramétrique des moindres carrés.

2.10.7 L'ultramétrique sous-dominante

Étant donné une dissimilarité d, la sous-dominante δ_d de d est la borne supérieure de l'ensemble des ultramétriques inférieures à d. Autrement dit, pour tout w_1 et w_2 dans Ω:

$$\delta_d(w_1,w_2) = \text{Sup}\{\delta(w_1,w_2): \delta \in U, \delta \le d\}.$$

Cet indice existe bien car l'ensemble des valeurs $\delta(w_1,w_2)$ est majorée par $d(w_2,w_2)$.

On montre que δ_d est une ultramétrique. δ_d étant définie comme une borne supérieure, elle est unique et appelée l'ultramétrique sous-dominante.

La proposition suivante permet de préciser le problème d'optimisation dont la sous-dominante est solution.

Soit Δ une mesure de ressemblance entre indices de dissimilarité telle que:

$$\Delta(d,\delta) = \left(\sum_{\substack{w_1 \in \Omega \\ w_2 \in \Omega}} (d(w_1,w_2)-\delta(w_1,w_2))^{\alpha} \right)^{1/\alpha}$$

avec $\alpha>0$.

Parmi les ultramétriques inférieures à d, l'ultramétrique la plus proche de d au sens de Δ est la sous-dominante. Autrement dit:

$$\Delta(d,\delta_d) = \text{Min } \{\Delta(d,\delta): \delta \in U, \delta \leq d\}$$

où U est l'ensemble des ultramétriques définies sur Ω.

On peut construire l'ultramétrique sous-dominante de plusieurs façons. On peut montrer que l'algorithme de la classification ascendante hiérarchique utilisant l'indice d'agrégation du lien minimum détermine la hiérarchie dont l'ultramétrique induite est la sous-dominante. On peut aussi la construire à partir de l'arbre de longueur minimum que nous allons présenter maintenant.

Définition d'un arbre de longueur minimum

On considère le graphe valué suivant $(\Omega, \Omega\times\Omega)$ où les sommets sont les éléments de Ω et les arêtes sont tous les couples de Ω. La valuation associée à chaque arête est la mesure de dissimilarité entre les deux sommets de cette arête. Un arbre est un graphe connexe sans cycle. Un arbre de longueur minimum sur Ω est un arbre de Ω dont la somme des valuations de ses arêtes est minimale.

Équivalence entre hiérarchie du lien minimum, arbre de longueur minimum et ultramétrique sous dominante

En supprimant successivement les arêtes les plus grandes de l'arbre de longueur minimum, on peut montrer que les parties connexes obtenues constituent la hiérarchie du lien minimum d'où l'équivalence entre les trois notions. Remarquons que la distance ultramétrique entre deux éléments de W est la longueur de la plus grande arête de la chaîne qui les relie.

Effet de chaîne: On doit noter que l'ultramétrique sous-dominante est sujette à l'effet de chaîne: il suffit qu'il existe entre deux points même très éloignés, une suite de points dont chaque point est très proche du suivant pour que ces deux points soient considérés comme très proches.

2.10.8 Ultramétrique sur-dominante

On pourrait définir l'ultramétrique sur-dominante de façon analogue:

$$\delta_s(w_1,w_2) = \text{Inf}\{\delta(w_1,w_2): \delta \in U, \delta \geq d\}$$

Cet indice existe bien car l'ensemble des valeurs $\delta(w_1,w_2)$ est minorée par $d(w_1,w_2)$. Mais, il se trouve que δ_s n'est pas forcément une ultramétrique.

Exemple: Soit l'indice de dissimilarité d suivant

	a	b	c
a	0	1	2
b		0	1
c			0

Les hiérarchies H_1 et H_2 induisent deux ultramétriques w_1 et w_2.

$$\delta_1:\ \begin{array}{c|ccc} & a & b & c \\ a & 0 & 1 & 2 \\ b & & 0 & 2 \\ c & & & 0 \end{array} \qquad \delta_2:\ \begin{array}{c|ccc} & a & b & c \\ a & 0 & 2 & 2 \\ b & & 0 & 1 \\ c & & & 0 \end{array}$$

δ_1 et δ_2 sont des ultramétriques supérieurs à d et leur enveloppe inférieure est d elle-même, c'est donc la borne inférieure δ_s et ce n'est pas une ultramétrique.

Toutefois, on définit la notion d'ultramétrique sur-dominante de la manière suivante:

$$\delta_s = \operatorname*{Min}_{\substack{\delta > d \\ \delta \in U}} \Delta(\delta,d)$$

où

$$\Delta(\delta,d) = \sum_{(w,w') \in \Omega \times \Omega} |\delta(w,w') - d(w,w')|$$

Cette définition ne donne pas l'unicité.

Notons que les hiérarchies du lien maximum qui fournissent une ultramétrique sur-dominante sont moins utilisées que celles du lien minimum et celles utilisant l'inertie que nous allons voir maintenant.

2.10.9 Hiérarchies basées sur l'inertie

On se place ici dans le cadre de l'optimisation des moindres carrés:

$$\Delta(d,\delta) = \sum_{w_1,w_2 \in \Omega} (d(w_1,w_2) - \delta(w_1,w_2))^2$$

CHANDON, LEMAIRE et POUGET (1980) ont proposé un algorithme permettant de donner une solution à ce problème. Malheureusement cet algorithme est inapplicable à de grands tableaux de données. La CAH utilisant l'indice d'agrégation de la moyenne des distances (SOKAL et MICHENER 1958)

$$\delta_u(h_1,h_2) = \frac{1}{p(h_1)p(h_2)} \sum_{\substack{w_i \in h_1 \\ w_j \in h_2}} d(w_i,w_j)$$

permet d'obtenir une solution approchée à ce problème d'optimisation. Un autre algorithme (WARD 1963) qui agrège à chaque itération deux classes de telle sorte que la partition obtenue ait une inertie intraclasse minimale, est beaucoup utilisée. Du fait qu'elle repose sur un critère d'inertie, on l'utilise souvent conjointement avec des analyses factorielles (BENZÉCRI 1973). On doit noter (cf. PERRUCHET 1983) que cette méthode à tendance à fournir des classes sphériques comme l'algorithme des centres mobiles.

D'autres approches non géométriques ont été étudiées. On doit citer l'algorithme de la vraisemblance du lien (AVL) (LERMAN 1981). Il s'agit d'une CAH conçue dans un cadre probabiliste. L'indice d'agrégation proposé mesure l'écart des données à une hypothèse simple d'absence de lien entre les objets à agréger. Ce cadre probabiliste s'applique à la plupart des tableaux de données recensés dans ce livre.

Conclusion

Nous avons vu des méthodes de classification de nature très différente qui fournissent des partitions ou des hiérarchies. Quand faut-il utiliser l'une plutôt que l'autre? Par construction, une hiérarchie définit une suite de partitions. C'est donc une structure plus riche mais plus compliquée qu'une partition. La plupart du temps l'utilisateur ne retient que quelques partitions d'une hiérarchie. Pourquoi ne pas rechercher directement une partition? L'examen de l'arbre hiérarchique est un moyen de déterminer un nombre de classes raisonnable, problème toujours difficile en classification. Pour les méthodes de partitionnement le nombre de classe doit être fixé a priori.

- Pour un nombre de classe donné, une hiérarchie fournit une seule partition ce qui n'est pas le cas pour les méthodes de partitionnement.
- Dans le cas de l'inertie, le critère utilisé par les méthodes de partitionnement vise à trouver directement la meilleure partition. Ce n'est pas le cas pour les hiérarchies pour lesquelles la partition retenue est un sous-produit: l'optimisation est seulement locale, elle se dégrade au fur et à mesure des agrégations. Elle sera en général très importante pour de grands tableaux de données.

- De plus, pour ces grands tableaux de données, l'utilisation de méthodes hiérarchiques est très difficile aussi bien au niveau de la complexité de l'algorithme que celle de la lecture de l'arbre.

En fait, pour la classification de très grands tableaux, il existe une stratégie répandue et efficace qui utilise conjointement la classification hiérarchique et les Nuées Dynamiques. Esquis-sons-la rapidement (se reporter à (MOLLIÈRE 1985) pour un exposé plus détaillé):

- On effectue les Nuées Dynamiques avec un grand nombre de classes (de l'ordre de la centaine). Cette première phase va nous permettre de travailler sur des uonnées agrégées et ainsi de corriger le caractère local de la CAH.
- On effectue une CAH sur les centres des classes obtenues à l'étape précédente en utilisant l'indice de l'augmentation d'inertie.
- L'examen de l'arbre hiérarchique va permettre de déterminer un ou plusieurs nombres de classes raisonnables.
- On utilise alors les Nuées Dynamiques pour optimiser les partitions suggérées par la CAH à l'étape précédente.
- Une analyse poussée des partitions retenues pourra se faire à l'aide des outils d'analyse d'une multi-partition que nous avons présentés au paragraphe 2.0.

Bibliographie

BALL G.H., HALL D.J. (1965), *Isodata, a novel method of data analysis and pattern classification*. Technical Report Stanford Research Institute.

BENZÉCRI J.P. et collaborateurs (1973), *L'analyse des Données*. Dunod. Tome 1 : *La Taxinomie*. Tome 2 : *L'Analyse des Correspondances*.

BERTIN J. (1977), *La graphique et le traitement graphique de l'information*. Flammarion.

BERTRAND P. (1986), *Etude de la représentation pyramidale*. Thèse de 3ème cycle, Université Paris IX-Dauphine.

BONNER R.H. (1964), *On some clustering techniques*. IBM Journal Vol. 22.

BREIMAN L., FRIEDMAN J.H., OHLSEN R.A., STONE C.J. (1984), Classification *and Regression Trees*. Wadsworth.

BRUYNOOGHE M. (1978), *Classification ascendante hiérarchique de grands ensembles de données ; un algorithme rapide fondé sur la construction des voisinages réductibles*. Cahiers de l'Analyse des Données Vol. 3 n°1.

BRYANT P., WILLIAMSON J. (1978), *Asymptotic Behaviour of classification ML estimates*. Biometrika Vol. 65.

CAILLEZ F., PAGES J.P. (1976), *Introduction à l'Analyse des Données*. S.M.A.S.H.

CAZES P. (1985), *L'Analyse Factorielle*. Polycopié ISUP.

CELEUX G. (1984), *Approximation rapide et interprétation d'une partition centrale pour les algorithmes de partitionnement*. Rapport de recherche INRIA n° 352.

CELEUX G. (1987), *Reconnaissance de mélanges de densités de probabilités et applications à la validation des résultats en classification*. Thèse d'état, Université Paris IX- Dauphine.

CELEUX G., DIEBOLT J. (1986), *L'algorithme SEM : un algorithme d'apprentissage probabiliste pour les mélanges de lois de probabilité*. R.S.A. Vol. 34 n° 2.

CELEUX G., LECHEVALLIER Y. (1982), *Méthodes de segmentation non paramétriques*. R.S.A. Vol. n° 30 n° 4.

CHANDON J.L., LEMAIRE J., POUGET J. (1980), *Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés*. RAIRO Vol. 14 n° 2.

CHARLES C. (1977), *Régression Typologique. Thèse 3ème cycle*, Université PARIS IX-Dauphine.

CHERNOFF H. (1970), *Metric considerations in the k-means method of cluster analysis*. Classification Society meetings. Ohio University.

COOPER L. (1972), *The transportation location problem*. Operat. Research Vol. 20 n° 1.

DEMPSTER A., LAIRD N., RUBIN D. (1977), *Maximum likelihood from incomplete data via the EM algorithm*. JRSS, B39.

DE RAHM C. (1980), *La classification hiérarchique ascendante selon la méthode des voisins réciproques*. Cahiers de l'Analyse des Données Vol. 5 n° 2.

DER MEGREDITCHIAN G. (1982), *Le traitement statistique des données multidimensionnelles*. EERM-ENM. Toulouse.

DIDAY E. (1972), *Optimisation en classification automatique et reconnaissance des formes*. Note Scient. IRIA n° 6.

DIDAY E. (1972b), *Nouveaux concepts et nouvelles méthodes en classification automatique*. Thèse d'état, Université Paris VI.

DIDAY E. et collaborateurs (1980), *Optimisation en classification automatique*. INRIA.

DIDAY E. (1986), *Une représentation visuelle des classes empiétantes : les pyramides*. RAIRO, APII Vol. 20 n° 5.

DIDAY E., LEMAIRE J., POUGET J., TESTU F. (1982), *Eléments d'analyse de données*. Dunod.

DUNSTAN J., WELSH D.J. (1973), *A greedy algorithm for solving a certain class of linear programms*. Mathem. Programming Vol. 5.

EVERITT B., HAND D. (1981), *Finite mixtures distributions*. Chapman and Hall.

FALGUEROLLES (DE) A. (1977), *Classification automatique : un critère et des algorithmes d'échanges*. Séminaire IRIA.

FISHER R.A. (1936), *The use of multiple measurements in taxonomic problems*. Eugenics Vol. 7.

FISHER W.D. (1958), *On grouping for maximum homogeneity*. JASA Vol. 53.

FORGY E.W. (1965), *Cluster analysis of multivariate data : efficiency versus interpretability of classifications*. Biometrics Vol. 21.

FRIEDMAN J.H. (1977), *A recursive partitioning decision rule for non parametric classification*. IEEE Trans. on Comp. C 26, 4.

FRIEDMAN H.P., RUBIN J. (1967), *On some invariant criteria for grouping data*. JASA Vol. 62.

GARFINKEL R., NEMHAUSER G.L. (1969), *The set-partitioning problem set covering with equality constraints*. Operat. Research Vol. 17.

GOVAERT G. (1975), *Classification automatique et distances adaptatives*. Thèse 3ème cycle, Université Paris VI.

GOVAERT G. (1983), *Classification croisée*. Thèse d'état, Université Paris VI.

GROSBRAS J.M. (1987), *Méthodes statistiques des sondages*. Economica.

HAND D.J. (1981), *Discrimination and Classification*. Wiley.

HARTIGAN J. (1975), *Clustering algorithms*. Wiley.

HUBERT L. (1974), *Some applications of graph theory to clustering*. Psychometrika Vol. 39.

JAMBU M., LEBEAUX M.O. (1978), *Classification automatique pour l'analyse des données*. Dunod.

JENSEN R.E. (1969), *A dynamic programming algorithm for cluster analysis*. Operat. Research Vol. 12.

LANCE G.N., WILLIAMS W.T. (1967), *A general theory of classification sorting strategies, I : hierarchical systems*. Computer Journal Vol. 9 n° 4.

LEBART L., MORINEAU A., TABARD N. (1977), *Techniques de la description statistique*. Dunod.

LECHEVALLIER Y. (1974), *Optimisation de quelques critères en classification automatique*. Thèse 3ème cycle, Université Paris VI.

LECHEVALLIER Y. (1979), *Classification optimale sous contrainte d'ordre total*. Rapport de recherche IRIA n° 200.

LEMOINE Y. (1979), *Classification et discrimination - Analyse discriminante typologique et applications*. Thèse 3ème cycle, Université de Metz.

LERMAN I.C. (1981), *Classification et analyse ordinale des données*. Dunod.

MAC QUEEN J.B. (1967), *Some methods for classification and analysis of multivariate observations*. Proc. of the 5th Berkeley Symposium on Math. Statistics and Probability. Vol. 1.

MARCOTORCHINO F., MICHAUD P. (1982), *Agrégation de similarités en classification automatique*. R.S.A. Vol. 30 n° 2.

MARRIOTT F. (1975), *Separating mixtures of normal distributions*. Biometrika Vol. 31.

MEUNIER P. (1986), *Segmentation et typologie*. Thèse, Université de Namur.

MINOUX M. (1977), *Algorithmes gloutons et algorithmes gloutons accélérés pour la résolution de grands problèmes combinatoires*. Bulletin EDF, Série C, n° 1.

MODULAD (1987), *Bibliothèque Fortran pour l'analyse des données*. INRIA.

MOLLIERE J.L. (1985), *What is the real number of clusters ?* 9th metting of the German Classification Society.

OK Y. (1975), *Analyse factorielle typologique et lissage typologique*. Thèse 3ème cycle, Université Paris VI.

PERRUCHET C. (1983), *Justification statistique de la classification ascendante hiérarchique suivant la variance*. Proc. des 3èmes journées internationales Analyse des Données et Informatique. North Holland.

PRIM R.C. (1957), *Shortest connection network and some generalizations*. Bell System Tech. Journal Vol. 36.

RALAMBONDRAINY H. (1986), *Contribution à l'analyse des données, I : Etude des tableaux n-aires, II : Le système SICLA*. Thèse d'état, Université Paris IX-Dauphine.

REDNER R., WALKER H. (1984), *Mixture densities, maximum likelihood and the EM algorithm*. SIAM Rev. Vol. 26.

RÉGNIER S. (1965), *Sur quelques aspects mathématiques des problèmes de classification automatique*. ICC Bulletin 4, et, réédité dans Math Sci. Hum. Vol. 82 (1983).

ROMEDER J.M. (1976), *Méthodes et Programmes d'analyse discriminante*. Dunod.

ROUX M. (1985), *Algorithmes de Classification*. Masson.

SALAS-ALVES D. (1983), *Structures des données récursives, application à la recherche de plus proches voisins et à la classification*. Thèse 3ème cycle, Université Paris VI.

SCHROEDER A. (1976), *Analyse d'un mélange de distributions de probabilité de même type*. R.S.A. Vol. 24 n° 1.

SCOTT A.J. (1969), *On the optimal partitioning of spatially distribued point sets*. Studies in Regional Science. ed : A.J. Scott.

SCOTT A., SYMONS M. (1971), *Clustering methods based on likelihood ratio criteria*. Biometrics Vol. 27.

SHLEZINGER I. (1968), *An algorithm for solving the self-organization problem*. Cybernetics n° 2.

SOKAL R.R., MICHENER C.D. (1958), *A statistical method fo evaluating systematic relationships*. Univ. Kansas Sci. Bull. Vol. 38.

SPAD (1985), *Système portable pour l'analyse des données* (Lebart L., Morineau A.). CESIA.

SYMONS M. (1981), *Clustering criteria and multivariate normal mixtures*. Biometrics Vol. 37.

THORNDIKE R.L. (1953), *Who Belongs in the family ?* Psychometrika Vol 18.

TOMASSONE R., DAUDIN J.J., DANZART M., MASSON J.P., (1987), *Discrimination et Classification*. Masson.

WARD J.H. (1963), *Hierarchical grouping to optimize an objective function*. JASA Vol. 58.

L'ANALYSE DES CORRESPONDANCES ET LE TRAITEMENT DES DONNÉES D'ENQUÊTES

Ludovic Lebart
CNRS, Paris

À la base du dépouillement d'enquête traditionnel, on trouve des techniques relativement frustes, éprouvées, faciles à mettre en oeuvre, faciles à interpréter: les tris, les tableaux croisés, c'est-à-dire en définitive des calculs de pourcentages et de moyennes.

On calculera par exemple les pourcentages des individus ayant choisi chaque item de réponse à une question fermée (qui constituent les *modalités* d'une *variable nominale*). Ces pourcentages seront calculés par rapport à l'échantillon global, mais aussi par rapport à des sous-échantillons, qui pourront d'ailleurs correspondre aux items d'une autre question. On dit que l'on a alors croisé les deux questions. Les variables *numériques* ou *quantitatives* (âge, nombre d'enfants, revenu, etc...) donnent lieu à des calculs de moyennes, qui peuvent être ventilées selon les catégories d'une ou de plusieurs variables nominales, ou à des calculs de corrélations, mesurant l'intensité du lien entre deux variables numériques. Plus rarement, des méthodes statistiques plus élaborées viennent compléter ces premiers calculs: régressions, analyses de la variance ou de la covariance, modèles log-linéaires.

L'utilisation des techniques d'analyse des données (ou encore: analyses descriptives multidimensionnelles) modifie profondément les premières phases du traitement des données d'enquête: ces techniques ne sont pas des compléments raffinés intervenant à la suite des méthodes traditionnelles. L'analyse des données va en fait révolutionner l'enchaînement des tâches, et définir une méthodologie et des concepts nouveaux.

La cohérence globale des données recueillies lors d'une enquête peut maintenant être éprouvée de façon rapide et systématique, des panoramas globaux peuvent être dressés, permettant de critiquer l'information, mais aussi d'orienter la suite des traitements, de choisir les "recodages" (transformations de variables par agrégation d'items, changement d'échelle, création de variables composites, etc...) et les tableaux croisés les plus pertinents. Les typologies (classification des individus en prenant en compte simultanément plusieurs réponses ou plusieurs caractéristiques de base), les outils de visualisation (plans factoriels) permettent de jeter un regard plus "macroscopique" sur l'information: l'utilisateur n'est plus "le nez sur les chiffres"...

Ces opérations, intervenant au début de la chaîne de traitement, permettent de piloter la suite du dépouillement de l'information. Non seulement le choix des modèles n'est plus fait de façon aveugle en fonction des hypothèses de base, mais ces hypothèses pourront souvent être critiquées, d'autres hypothèses pourront être suggérées.

Le présent texte, qui correspond pour l'essentiel au chapitre 3 de l'ouvrage *Analyse des Données Textuelles* (par L. LEBART et A. SALEM, Dunod, 1988), rappelle brièvement les principes de base et les modalités pratiques d'utilisations de ces méthodes, avant de mettre en relief leurs contributions au renouvellement des techniques de dépouillement d'enquêtes socio-économiques.

École d'été sur l'analyse des données, juin 1988.
UQAM, SCAD-CIRADE.

1 Principes de base des techniques d'analyse des données

Le principe commun à toutes les méthodes de statistique descriptive multidimensionnelle est schématiquement le suivant:

Chaque dimension d'un tableau rectangulaire de données numériques permet de définir des distances (ou des proximités) entre les éléments de l'autre dimension: ainsi, l'ensemble des colonnes (qui peuvent être des variables, des attributs) permet de définir à l'aide de formules appropriées des distances entre lignes (qui peuvent être des individus, des observations). De la même façon, l'ensemble des lignes permet de calculer des distances entre colonnes. On obtient ainsi des tableaux de distances, qui sont associées à des représentations géométriques complexes des similitudes existant entre les lignes et entre les colonnes des tableaux rectangulaires à décrire.

Le problème est de rendre assimilable et accessible à l'intuition ces représentations, au prix d'une perte d'information de base qui doit rester la plus petite possible. Brièvement, on peut dire qu'il existe deux famille de méthodes qui permettent d'effectuer ces réductions:

- *Les méthodes factorielles*, fondées sur l'algèbre linéaire, qui produisent des représentations graphiques sur lesquelles les proximités géométriques usuelles entre points-lignes et entre points-colonnes traduisent les associations statistiques entre lignes et entre colonnes. C'est à cette famille de méthodes qu'appartient l'analyse des correspondances.

- *Les méthodes de classification* qui opèrent des regroupements en classes (ou en familles de classes hiérarchisées) des lignes et des colonnes.

Ces deux familles de méthodes peuvent d'ailleurs être utilisées avec profit sur les mêmes tableaux de données, et se compléter utilement.

Il va de soi que les règles d'interprétation des représentations obtenues à l'issue de ces techniques de réduction n'ont pas la simplicité de celles de la statistique descriptive élémentaire. L'interprétation des histogrammes, les diagrammes en bâtons, les graphiques de séries chronologiques ne nécessitent qu'un apprentissage rudimentaire; alors que dans le cas de l'analyse des correspondances, par exemple, il sera nécessaire de connaître des règles de lectures des résultats plus contraignantes que ne le laisse croire le caractère souvent suggestif des représentations obtenues. Il faudra au lecteur non seulement un apprentissage, mais aussi une véritable expérience clinique (on veut dire par là une expérience variée et difficilement formalisable: chaque tableau de données est en quelque sorte un cas particulier).

Pour bien faire assimiler le fonctionnement d'une méthode, il est nécessaire de raisonner sur un exemple de dimension réduite parfaitement maîtrisable par l'utilisateur... mais les techniques d'analyse exploratoire multidimensionnelle ne présentent de l'intérêt que parce qu'elles permettent de traiter de très grands ensembles de données. Il est donc difficile de démontrer l'efficacité de la méthode en même temps que ses principes de fonctionnement.

Nous allons donc travailler sur des tableaux de dimensions très réduites qui illustreront les principes, mais ne permettront pas d'apprécier pleinement l'efficacité des outils utilisés.

2 L'analyse des correspondances

2.1 Bref historique

L'analyse des correspondances est une technique de description des tables de contingences (ou encore: tableaux croisés) et de certains tableaux binaires (tableaux dits de "présence-ab-

sence). Cette description se fait essentiellement sous forme de représentation graphique des associations entre lignes et entre colonnes.

Présentée et étudiée de façon systématique comme une technique "versatile" d'analyse exploratoire de données multidimensionnelles par JEAN-PAUL BENZECRI (1973) dans un ouvrage qui reprend les travaux de son laboratoire au cours des dix années précédentes, l'analyse des correspondances s'est trouvé depuis de nombreux antécédents et précurseurs, et a donné lieu à des travaux dispersés et indépendants les uns des autres. On peut citer sans être exhaustif les noms de [GUTTMAN 1941] et de [HAYASHI 1956] qui parlent tout deux de "méthodes de quantification", [NISHISATO 1980] qui parle de "dual scaling", [GIFI 1981] qui parle d'"homogeneity analysis". On trouvera dans [NISHISATO 1980, 1986] et dans [TENENHAUS et al. 1985] plusieurs centaines de références sur le sujet.

Tableau 1: Croisement "Formes-Niveau de Diplôme"
(Effectifs bruts)

	Sans Diplôme	CEP	BEPC	Bac	Univ	TOTAL
Peur	25	45	38	38	13	159
Santé	18	27	20	19	9	93
Avenir	53	90	78	75	22	318
Argent	51	64	32	29	17	193
Emploi	12	35	19	6	7	79
Guerre	4	7	7	6	2	26
Chômage	71	111	50	40	11	283
Travail	35	61	29	14	12	151
Égoïsme	21	37	14	26	9	107
Maladie	21	37	14	26	9	107
Finances	10	7	7	3	1	28
Logement	8	22	7	10	5	52
Difficile	7	11	4	3	2	27
Économique	7	13	12	11	11	54
Financières	21	32	42	47	30	172
Conjoncture	1	7	5	5	4	22
TOTAL	365	606	378	358	164	1871

2.2 Principes de base de l'Analyse des Correspondances exposés à partir d'un exemple numérique simple

Le tableau 1 que l'on va s'efforcer de décrire avec l'aide de l'analyse des correspondances est un tableau de fréquences croisant, en ligne, 16 mots (il s'agit ici de formes graphiques, et pas exactement de mots) utilisés dans leurs réponses à une question ouverte par 2000 personnes, et en colonne, 5 niveaux de diplômes[1] de ces mêmes personnes. Le libellé de la question ouverte était le suivant: "*Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant?*"

[1] *Sans diplôme, CEP (Certificat d'études primaires), BEPC (Brevet d'études du premier cycle ou équivalent), Bac (Baccalauréat ou équivalent), Univ. (Université, grandes écoles ou équivalent).*

Notons que l'unité statistique retenue ici n'est pas l'individu, mais l'occurrence d'une forme graphique. En effet, les colonnes constituent bien une partition de l'ensemble des personnes interrogées, mais pas les lignes: un répondant peut utiliser dans ces réponses plusieurs formes de la liste retenue ici.

On lit ce tableau de la manière suivante: la forme "*argent*", par exemple, a été utilisé 51 fois dans leurs réponses libres par les personnes sans diplôme.

Les totaux de chaque ligne représentent les nombres d'occurrences de chaque forme alors que les totaux de chaque colonne représentent les nombres totaux de formes (de la liste) utilisées par les diverses catégories de diplômes.

Le tableau 2 représente les profils-lignes, exprimés ici en pourcentages: ils sont obtenus en divisant chaque élément du tableau par la somme de la ligne correspondante: le profil-ligne de la forme "*peur*", par exemple, est obtenu en divisant chaque terme de la ligne par 159. Pour des questions de lisibilité, les nombres obtenus sont multipliés par 100.

La comparaison de deux profils-lignes va nous renseigner sur la façon dont les deux formes correspondantes s'associent aux diplômes. Cette comparaison est plus difficile à partir du seul tableau 1, car les fréquences des formes sont très variables.

Lorsque deux formes seront représentées par des points voisins en analyse des correspondances, cela signifiera que les profils-lignes correspondants sont voisins.

Tableau 2: Les profils-lignes du tableau 1

	Sans Diplôme	CEP	BEPC	Bac	Univ	TOTAL
Peur	15.7	28.3	23.9	23.9	8.2	100.0
Santé	19.2	29.0	21.5	20.4	9.7	100.0
Avenir	16.7	28.3	24.5	23.6	6.9	100.0
Argent	26.4	33.2	16.6	15.0	8.8	100.0
Emploi	15.2	44.3	24.1	7.6	8.8	100.0
Guerre	15.4	26.9	26.9	23.1	7.7	100.0
Chômage	25.1	39.2	17.7	14.1	3.4	100.0
Travail	23.2	40.4	19.2	9.3	7.9	100.0
Égoïsme	19.6	34.6	13.1	24.3	8.4	100.0
Maladie	19.6	34.6	13.1	24.3	8.4	100.0
Finances	35.7	25.0	25.0	10.7	3.6	100.0
Logement	15.3	42.3	13.5	19.2	9.6	100.0
Difficile	25.9	40.7	14.8	11.1	7.4	100.0
Économique	13.0	24.1	22.2	20.4	20.4	100.0
Financières	12.2	18.6	24.4	27.3	17.4	100.0
Conjoncture	4.5	31.8	22.7	22.7	18.3	100.0
MARGE	19.5	32.4	20.2	19.1	8.8	100.0

Le tableau 3 va maintenant représenter les profils-colonnes: ceux-ci sont obtenus de façon tout à fait analogue en divisant les éléments de chaque colonne par leur somme, et en multipliant par 100 le résultat obtenu. La comparaison de deux profils-colonnes nous renseigne sur les proxi-

mités existant entre les différentes catégories de diplômes vis-à-vis du vocabulaire employé.

Le voisinage de deux points représentant les catégories de diplômes en analyse des correspondances traduira une similitude de profils-colonnes.

L'analyse des correspondances va donc décrire simultanément les similitudes de profils-lignes et de profils-colonnes, et fournir une représentation schématique des informations contenues dans les tableaux 2 et 3.

La figure 1 (plan factoriel engendré par les deux premiers axes de l'analyse des correspondances du tableau 1) donne une représentation visuelle des associations entre lignes et colonnes de ce tableau.

Pour une personne qui en maîtrise bien les règles de lecture, c'est un procédé très rapide d'assimilation de l'information. On ne peut cependant pas attendre de la description d'une table dont les dimensions sont aussi modestes des grandes surprises ou des révélations.

Tableau 3: Profils-colonnes du tableau 1

	Sans Diplôme	CEP	BEPC	Bac	Univ	MARGE
Peur	6.8	7.4	10.0	10.6	7.9	8.5
Santé	4.9	4.4	5.3	5.3	5.5	4.9
Avenir	14.5	14.8	20.6	20.9	13.4	17.0
Argent	13.9	10.6	8.5	8.1	10.4	10.3
Emploi	3.3	5.8	5.0	1.7	4.3	4.2
Guerre	1.1	1.2	1.8	1.7	1.2	1.4
Chômage	19.4	18.3	13.2	11.2	6.7	15.1
Travail	9.6	10.1	7.7	3.9	7.3	8.1
Égoïsme	5.7	6.1	3.7	7.3	5.5	5.7
Maladie	5.7	6.1	3.7	7.3	5.5	5.7
Finances	2.7	1.2	1.8	0.8	0.6	1.5
Logement	2.2	3.6	1.8	2.8	3.0	2.8
Difficile	1.9	1.8	1.1	0.8	1.2	1.5
Economique	1.9	2.1	3.2	3.1	6.7	2.9
Financières	5.7	5.3	11.1	13.1	18.2	9.2
Conjoncture	0.3	1.2	1.3	1.4	2.4	1.2
TOTAL	100.0	100.0	100.0	100.0	100.0	100.0

L'analyse des correspondances a ici une fonction purement descriptive: la consultation des résultats est plus agréable et confortable.

Montrons, à propos de cet exemple, la simplicité des règles d'interprétation et des principes de la méthode.

On notera $f_{i,j}$ le terme général de la table des fréquences dont les éléments sont préalablement divisés par l'effectif total k (k=1871 pour le Tableau 1). Cette table possède n lignes et p colonnes (ici, n=16 et p=5). Selon les notations usuelles, $f_{i.}$ et $f_{.j}$ désignent les sommes des lignes et des colonnes de cette table. Le profil de la ligne i est l'ensemble des p valeurs:

$$\frac{f_{i,j}}{f_{i.}}, \quad j=1, \ldots, p.$$

Le profil de la colonne j est l'ensemble des n valeurs:

$$\frac{f_{ij}}{f_{.j}}, \quad i=1, \ldots, n.$$

Les éléments des profils, multipliés par 100, ne sont autres que les pourcentages-lignes et les pourcentages-colonnes représentés dans les tableaux 2 et 3.

Comment lire la Figure 1?

Si deux points lignes i et i' ont des profils identiques ou voisins, ils seront confondus ou proches sur chacun des axes factoriels. De façon tout à fait symétrique, si deux points colonnes j et j' ont des profils identiques ou voisins, ils seront confondus ou proches.

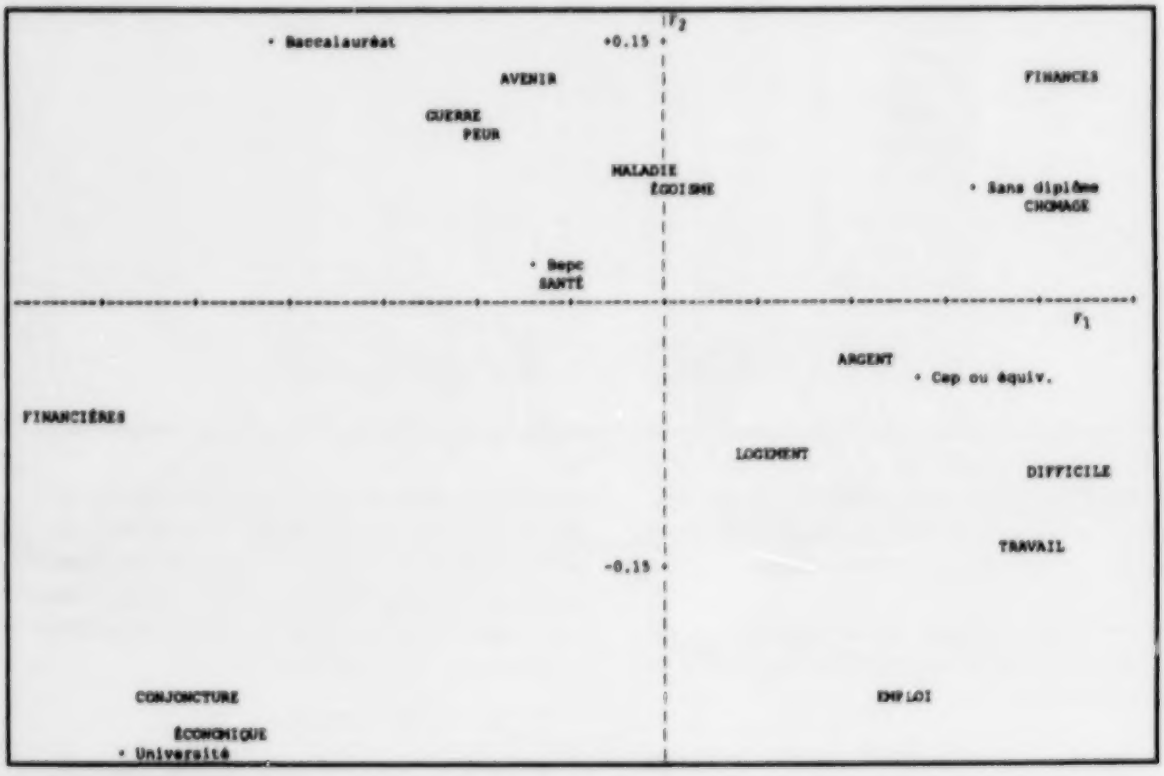
L'origine des axes correspond aux profils moyens (marges de la table de fréquences). Ainsi un point-colonne comme "Bepc" a un profil voisin de celui de la colonne Marge du tableau 3. De la même façon, un point ligne comme *"santé"* a un profil voisin de la ligne Marge du tableau 2.

Le profil-ligne moyen a pour composantes $f_{.j}$, $j=1,..p$.

Le profil-colonne moyen a pour composantes $f_{i.}$, $i=1,..n$.

Figure 1

Associations entre formes et diplômes
(Réponses libres sur "les raisons pouvant faire hésiter un couple à avoir un enfant")
Analyse des correspondances du tableau 1



Les points occupant des positions périphériques auront donc les profils les plus différents du profil moyen, donc les plus typés. Tels sont les cas des mots comme *"finances"* et *"financières"*, diamétralement opposés, car correspondant respectivement aux formulations populaires et instruites d'une même préoccupation.

Que signifient les proximités?

Il reste cependant à définir ce que l'on appelle "proche", c'est à dire à expliciter la façon dont sont calculées les distances dont des graphiques tels que la figure 1 fournissent des approximations planes.

Les profils qui sont des suites de n ou p nombres, selon qu'il s'agit de lignes ou de colonnes, permettent de définir des points dans des espaces à n ou p dimensions (plus précisément à n-1 et p-1 dimensions, car les sommes de leur composantes valent 1).

Les distances entre points seront donc définies dans ces espaces de dimensions élevées. La phase d'analyse proprement dite consistera à réduire ces dimensions de façon à permettre une représentation visuelle, tout en déformant le moins possible les distances.

La distance entre deux points-lignes i et i' est donnée par la formule:

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{i,j}}{f_{i.}} - \frac{f_{i',j}}{f_{i'.}} \right)^2 .$$

De la même façon, la distance entre deux points-colonnes j et j' est donnée par:

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{i,j}}{f_{.j}} - \frac{f_{i,j'}}{f_{.j'}} \right)^2 .$$

Cette distance, appelée distance du χ^2, ressemble beaucoup à la distance euclidienne usuelle (somme des carrés des différences entre composantes des profils) à ceci près qu'une pondération intervient (inverse de la fréquence correspondant à chaque terme).

Cette distance a une propriété remarquable, dite d'*équivalence distributionnelle*. Disons en bref que cette propriété assure une invariance des distances entre lignes (resp. colonnes) lorsque l'on agrège deux colonnes (resp. lignes) ayant des profils identiques. Ce principe d'invariance, assure une certaine stabilité des résultats vis à vis des nomenclatures: ainsi, deux classes de diplômes ayant même profil lexical peuvent être considérées comme distinctes ou non sans que cela modifie les représentations obtenues. Pour une justification plus argumentée de ces pondérations, cf. par exemple [BENZÉCRI 1973].

Pourquoi une représentation simultanée?

Nous savons maintenant interpréter la proximité de deux points-lignes ou de deux points-colonnes ainsi que leurs positions par rapport à l'origine des axes.

Mais la Figure 1 nous montre des points-lignes et des points-colonnes simultanément, et donc des proximités qu'il est bien tentant d' interpréter: que le point-ligne *"chômage"* soit au voisinage du point-colonne "Sans Diplôme" n'a peut-être rien de surprenant. Mais la proximité entre *"santé"* et "Bepc" est moins évidente. En fait, il est formellement déconseillé d'interpréter ces proximités croisées entre lignes et colonnes, car les deux points ne sont pas dans le même espace au départ: ce qui est licite, c'est d'interpréter la position d'un point-ligne par rapport à l'ensemble des points-colonnes (et vice-versa).

La principale justification de cette représentation simultanée est donnée par les formules dites de transitions, liant les coordonnées d'un point dans un espace (les lignes par exemple) à celles de tous les points de l'autre espace (les colonnes pour notre exemple).

Si ϕ_i désigne la coordonnée d'un point-ligne i sur l'axe horizontal de la figure 1 (premier axe factoriel), et si ψ_j désigne celle d'un point-colonne j sur ce même axe, on a le système de relations remarquablement symétriques:

$$\phi_i = \beta \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}} \psi_j \qquad (1)$$

$$\psi_j = \beta \sum_{i=1}^{n} \frac{f_{ij}}{f_{.j}} \phi_i \qquad (2)$$

Le coefficient β est un coefficient positif, supérieur à 1. C'est d'ailleurs le coefficient le plus proche de 1 qui soit compatible avec ces deux formules.

Sans ce coefficient, on voit que les points-lignes seraient des barycentres des points-colonnes, et réciproquement. Une telle représentation doublement barycentrique n'est pas possible, car la prise de barycentre a un effet contractant. Il faut donc un coefficient dilatateur β supérieur à 1.

On peut ainsi présenter l'analyse des correspondances comme la recherche des valeurs de ϕ_i et de ψ_j correspondant au plus petit coefficient dilatateur β (cf. [LEBART et al. (1979)]).

Les formules (1) et (2) sont également valables pour les coordonnées sur l'axe vertical de la figure 1, avec bien sûr des valeurs différentes pour les ϕ_i, ψ_j, et le coefficient β.

Autre propriété des coordonnées des points-lignes (et des points-colonnes) sur les axes factoriels: elles sont centrées, c'est à dire vérifient les relations:

$$\sum_{i=1}^{n} f_{i.}\phi_i = 0 \text{ et } \sum_{j=1}^{p} f_{.j}\psi_j = 0.$$

2.3 Validité de la représentation

Le tableau 4 donne les valeurs de deux séries de paramètres dont nous n'avons pas encore parlé: les valeurs propres, désignées par λ (comprises entre 0 et 1 en analyse des correspondances), qui égalent ici 0.05 pour le premier axe et 0.01 pour le second; les pourcentages de variance correspondant à ces valeurs propres qui ont ici les valeurs 65% et 15% pour ces deux mêmes axes.

Un autre paramètre, la trace t (somme de toutes les valeurs propres) vaut ici 0.076: il y a 4 valeurs propres non nulles. Alors que la trace représente l'inertie totale (ou variance totale) du nuage, les valeurs propres représentent les inerties (ou variances) correspondant à chaque axe.

Tableau 4: Valeurs propres et pourcentages de variance (ou d'inertie)

Numéro	*Valeur propre*	*Pourcentage*	*Pourcentage cumulé*
1	0.0496	65.44	65.44
2	0.0113	14.86	80.30
3	0.0082	10.87	91.18
4	0.0067	8.82	100.00

Propriétés de la trace t

Lors de l'analyse des correspondances d'une table de contingence (n,p), le produit ($k \times t$) de la trace t par l'effectif total k n'est autre que le classique Khi-2 (χ^2 de K. Pearson (avec $(n-1)\times(p-1)$ degrés de liberté) utilisé pour tester l'indépendance des lignes et des colonnes de la table.

Classiquement, ce paramètre est calculé par la formule:

$$\chi^2 = k \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

On a pour le tableau 1:

$$k\,t = 1871 \times 0.076 = 142.2$$

pour 60 degrés de libertés [$60 = (16-1) \times (5-1)$].

Bien entendu, l'hypothèse d'indépendance est rejetée sans qu'il soit nécessaire de consulter une table... et l'analyse des correspondances est précisément là pour nous aider à comprendre pourquoi cette hypothèse est rejetée.

Les valeurs propres

Pour un axe donné, il existe la relation suivante entre le coefficient β des relations de transition et la valeur propre:

$$\beta = \frac{1}{\lambda}$$

Une valeur propre proche de 1 assure donc une bonne représentation barycentrique le long de l'axe correspondant, ce qui n'est pas vraiment le cas pour cet exemple: si les points-diplômes étaient représentés comme de vrais barycentres des points-mots, ils seraient beaucoup moins excentrés... d'où le caractère peut-être trompeur, pour les non-initiés, de la représentation simultanée.

Tableau 5: *Principaux paramètres de l'analyse des correspondances du tableau 1. Les individus actifs sont les lignes du tableau 1 alors que les fréquences actives en sont les colonnes*

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS SUR LES AXES 1 A 4

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
PEUR	6.50	0.03	-0.14	0.09	-0.05	-0.06	0.00	3.3	5.0	2.6	2.5	0.0	0.61	0.24	0.08	0.06	0.00
SANTE	4.97	0.01	-0.07	0.01	-0.02	0.04	0.00	0.5	0.1	0.1	1.0	0.0	0.73	0.02	0.04	0.21	0.00
AVENIR	17.00	0.03	-0.11	0.12	-0.07	-0.04	0.00	4.0	21.6	10.9	3.8	0.0	0.36	0.44	0.16	0.05	0.00
ARGENT	10.32	0.04	0.14	-0.04	0.05	0.12	0.00	4.3	1.3	3.1	23.7	0.0	0.52	0.03	0.06	0.38	0.00
EMPLOI	4.22	0.13	0.17	-0.22	-0.16	-0.16	0.00	2.5	18.6	13.1	16.0	0.0	0.23	0.38	0.20	0.19	0.00
GUERRE	1.39	0.05	-0.15	0.10	-0.12	-0.04	0.00	0.6	1.2	2.6	0.3	0.0	0.47	0.19	0.31	0.03	0.00
CHOMAGE	15.13	0.07	0.27	0.04	0.00	0.00	0.00	21.9	2.5	0.0	0.0	0.0	0.97	0.03	0.00	0.00	0.00
TRAVAIL	9.07	0.08	0.26	-0.13	-0.05	0.00	0.00	9.5	13.0	2.4	0.0	0.0	0.74	0.23	0.03	0.00	0.00
EGOISME	5.72	0.04	0.00	0.06	0.19	-0.04	0.00	0.0	1.7	24.9	1.3	0.0	0.00	0.08	0.88	0.04	0.00
MALADIE	5.72	0.04	0.00	0.06	0.19	-0.04	0.00	0.0	1.7	24.9	1.3	0.0	0.00	0.08	0.88	0.04	0.00
FINANCES	1.50	0.23	0.27	0.12	-0.19	0.33	0.00	2.1	2.0	6.7	24.2	0.0	0.30	0.07	0.16	0.47	0.00
LOGEMENT	2.78	0.06	0.06	-0.09	0.15	-0.16	0.00	0.2	2.1	9.0	11.0	0.0	0.06	0.14	0.38	0.42	0.00
DIFFICILE	1.44	0.09	0.28	-0.09	0.06	0.04	0.00	2.3	1.1	0.5	0.3	0.0	0.86	0.09	0.03	0.02	0.00
ECONOMIQUE	2.89	0.20	-0.35	-0.26	0.02	0.09	0.00	7.3	17.4	0.1	2.6	0.0	0.43	0.34	0.00	0.03	0.00
FINANCIERES	9.19	0.22	-0.65	-0.07	0.01	0.07	0.00	38.2	3.9	0.0	6.9	0.0	0.96	0.02	0.00	0.02	0.00
CONJONCTURE	1.18	0.23	-0.37	-0.24	0.02	-0.17	0.00	3.2	6.1	0.0	5.3	0.0	0.41	0.26	0.00	0.13	0.00

FREQUENCES			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDEN - LIBELLE COURT	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
FREQUENCES ACTIVES:																	
SANS - SANS DIPLOME	19.51	0.07	0.23	0.04	0.02	0.14	0.00	20.6	3.2	1.2	55.4	0.0	0.71	0.03	0.01	0.25	0.00
- CEP	32.39	0.05	0.19	-0.05	0.03	-0.09	0.00	23.2	6.7	2.6	35.1	0.0	0.78	0.05	0.01	0.16	0.00
- BEPC	20.20	0.04	-0.11	0.02	-0.17	0.06	0.00	5.3	0.5	74.0	0.0	0.0	0.30	0.01	0.70	0.00	0.00
- BAC	19.13	0.10	-0.28	0.14	0.09	-0.03	0.00	26.9	34.4	17.4	2.1	0.0	0.71	0.21	0.08	0.01	0.00
UNIV - UNIVERSITE	8.77	0.22	-0.37	-0.27	0.07	0.07	0.00	23.9	55.2	4.9	7.5	0.0	0.63	0.33	0.02	0.03	0.00

Guide de lecture du tableau 5

La colonne "P.REL" (Poids relatifs) désigne les marges des lignes et des colonnes, qui figuraient déjà dans les tableaux 2 et 3 précédents.

La colonne "DISTO" contient les distances à l'origine des axes, c'est à dire les distances de chaque profil au profil moyen (ou: marge). On peut vérifier par exemple sur le tableau 2 que les mots "*finances*" et "*conjoncture*" ont des profils très différents de la marge.

Les deux premières "COORDONNEES" sont celles des points, dont la figure 1 fournit une représentation approchée.

Les "CONTRIBUTIONS" (ou encore: contributions absolues), dont les sommes en colonne valent 100, traduisent l'importance des différents éléments dans la construction de chaque axe.

Les "COSINUS CARRES" (ou encore: contributions relatives), dont les sommes en ligne valent 1, montrent l'importance des différents axes dans l'explication de chaque élément.

Tableau 6: 4 lignes supplémentaires (ou illustratives)

	Diplôme	Sans CEP	BEPC	Bac	Univ	TOTAL
Vivre	3	3	1	3	4	14
Monde	1	5	4	6	3	19
Confort	2	4	3	1	4	14
Mésentente	2	8	2	5	2	19

N.B. La racine carrée de la plus grande valeur propre peut également être interprétée comme le plus grand coefficient de corrélation existant entre les lignes et les colonnes de la table (corrélation canonique).

Ce coefficient se calcule de la façon suivante: supposons qu'à chacune des 5 catégories de diplôme correspondent une valeur numérique (5 valeurs différentes possibles), et qu'il en soit de même pour les mots utilisés (opération dite de codage ou de quantification). A chaque occurrence, on peut faire correspondre deux valeurs numériques (une pour les lignes, une autre pour les colonnes). On peut donc calculer un coefficient de corrélation entre ces deux ensembles de valeurs. Les 1871 occurrences peuvent être regroupées en 80 groupes (les 80 cases du tableau 1) correspondant à des couples de valeurs distinctes. La valeur maximale que peut atteindre ce coefficient de corrélation (qui est associée à un "codage optimal" des mots et des diplômes) est 1.

Les pourcentages de variance (ou d'inertie)

Ils mesurent la place de chaque valeur propre dans la trace. Il est assez exceptionnel que, comme c'est le cas ici, le premier plan factoriel "explique" 80% de la variance totale. En général, ces pourcentages sont au contraire une mesure pessimiste de la part d'information représentée. De nombreux contre-exemples montrent que des pourcentages médiocres correspondent parfois à des représentations qui rendent compte de façon satisfaisante de la structure des données.

Autres aides à l'interprétation

Deux séries de paramètres permettent d'interpréter avec plus de sûreté les résultats, en complétant l'information donnée par les coordonnées des éléments sur les axes factoriels:

Les contributions (ou contributions absolues), qui décrivent la part prise par un élément (ligne ou colonne) dans la construction d'un axe factoriel.

Les cosinus carrés (ou contributions relatives) qui mesurent la qualité de la représentation de chaque élément par les axes.

Le tableau 5 présente les valeurs de ces différents paramètres pour l'exemple présenté ici.

2.4 Variables actives et illustratives

L'analyse des correspondances permet de trouver des sous-espaces de représentation des proximités entre profils. Mais elle permet aussi de positionner dans ce sous-espace des lignes ou des colonnes supplémentaires du tableau de données. Une fois trouvés les axes factoriels Φ et Ψ et la valeur propre λ (et donc β), les formules de transition (1) et (2) ci-dessus peuvent être appliquées à des lignes ou des colonnes supplémentaires:

La formule (1) permet, à partir de ψ, β, de calculer, pour chaque axe factoriel, la coordonnée $\phi+$ d'une ligne supplémentaire.

La formule (2) permet de la même façon,

à partir de ϕ, β, de calculer la coordonnée $\psi+$ d'une colonne supplémentaire.

Il s'agit dans les deux cas de projections de points-profils sur les axes (et donc sur les plans..) factoriels.

On peut ainsi illustrer les plans factoriels par des informations supplémentaires n'ayant pas participé à la construction des plans, ce qui va avoir des conséquences très importantes au niveau de l'interprétation des résultats.

Les éléments ou variables servant à calculer les plans factoriels sont appelées éléments actifs ou
variables actives: ils doivent former un ensemble homogène pour que les distances entre individus ou observations s'interprètent facilement.

Tableau 7: 4 colonnes supplémentaires (ou illustratives)

	Autres Diplômes	- 30 ans	30 à 50 ans	+ de 50 ans
Peur	0	48	59	52
Santé	2	13	29	53
Avenir	0	115	117	86
Argent	2	59	66	70
Emploi	0	21	30	28
Guerre	0	7	6	13
Chômage	1	79	88	177
Travail	0	30	63	58
Égoïsme	2	14	34	61
Maladie	2	2	15	21
Finances	0	8	12	8
Logement	0	10	25	17
Difficile	0	2	17	18
Économique	0	18	19	17
Financière	11	62	72	39
Conjoncture	0	9	8	5
TOTAL	20	497	660	723

Les éléments (lignes ou colonnes) intervenant a posteriori sont appelés éléments illustratifs ou supplémentaires, ou variables illustratives (ou supplémentaires). Ils n'ont nul besoin de former un ensemble homogène dans la mesure où leur calcul est effectué a posteriori et séparément pour chacun d'eux.

Cette dichotomie entre variables actives et variables illustratives est fondamentale. Elle est du même ordre que la distinction que l'on établit entre variables endogènes (à expliquer) et exogènes (explicatives) dans les modèles de régression multiple. D'un point de vue géométrique, les deux situations sont d'ailleurs très similaires. Les variables exogènes engendrent un sous-espace sur lequel seront projetées les variables endogènes. Les variables actives engendrent aussi un sous-espace, que l'on va réduire pour le visualiser. C'est sur cet espace réduit que l'on va projeter les variables illustratives.

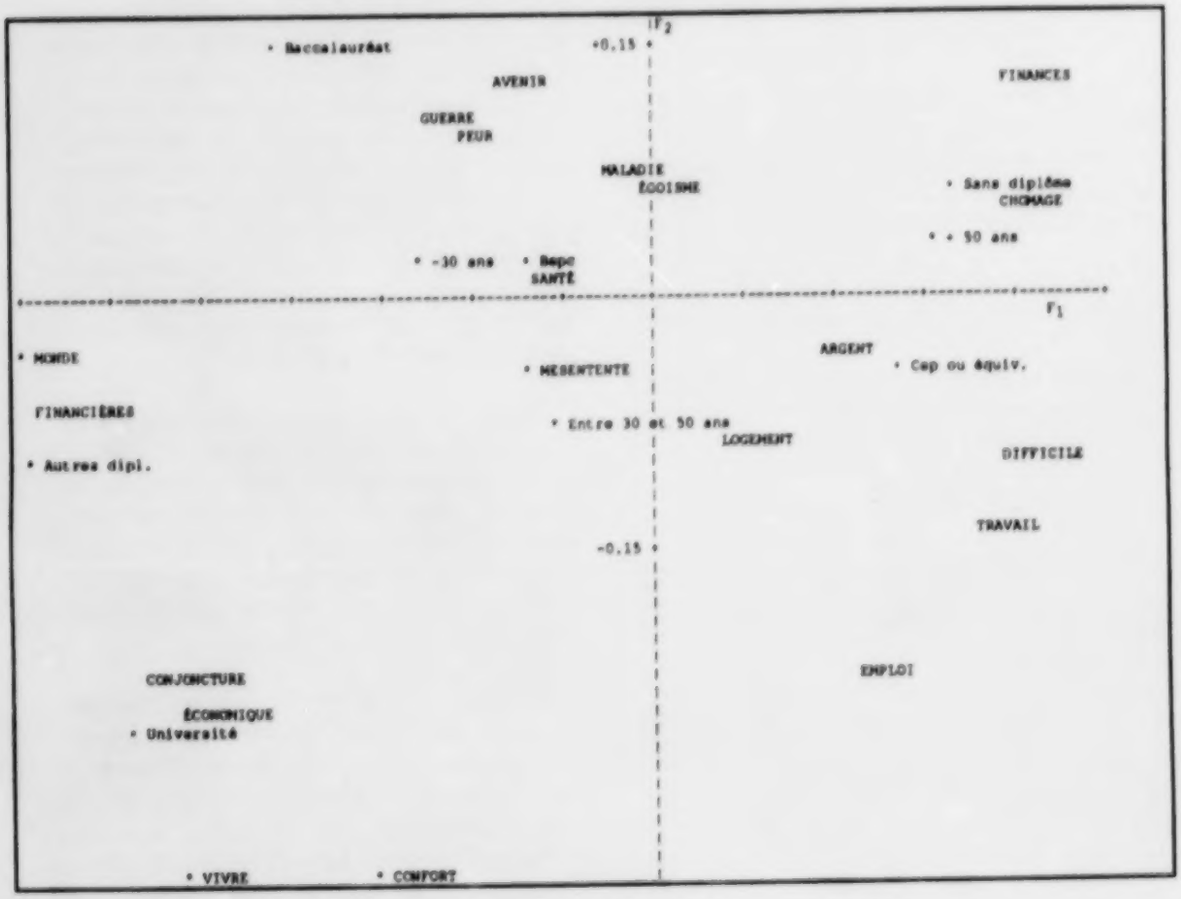
Ces quatre formes graphiques d'effectifs faibles dont on peut voir la ventilation sur le tableau 6 n'ont pas participé à l'analyse précédente. On désire cependant savoir comment elles se situent par rapport aux autres mots. Leurs profils-lignes peuvent être positionnés dans le même espace à 5 dimensions, et peuvent donc subir la même projection sur le plan de la figure 1.

La figure 2 nous montre que les trois premières formes appartiennent plutôt aux réponses des personnes dont le niveau de diplôme est élevé, alors que la quatrième, "*mésentente*", est moins caractéristique. Les paramètres plus techniques figurent dans le tableau 8. On ne s'étonnera pas de ne pas trouver de "contributions" dans ce tableau, puisque les éléments illustratifs ont par définition une contribution nulle: ils ne participent pas à la construction des axes, mais sont positionnés a posteriori.

La catégorie "Autres Diplômes" n'a pas été retenue comme colonne active en raison de l'extrême faiblesse de ses effectifs. Cependant, il nous a paru intéressant de la situer par rapport aux autres catégories de diplômes, en la projetant comme élément illustratif: La figure 2 nous montre que le point qui représente cette catégorie

Figure 2

Associations entre formes et diplômes (suite)
*Positionnement des éléments illustratifs (précédés du symbole *) sur la figure 1*



occupe une position intermédiaire entre le point "Bac"et le point "Université".

Les trois autres colonnes supplémentaires sont des catégories d'âge, qui n'ont pas été incluses dans l'ensemble des colonnes actives en raison de l'hétérogénéité des thèmes: l'interprétation des proximités entre lignes, donc entre mots, aurait été plus délicate: deux mots sont-ils proches à cause de leur répartition par diplômes ou par classes d'âges?

Il est de bonne discipline d'utiliser comme tableau actif des tableaux homogènes, ne décrivant les proximités que d'un point de vue.

Il est ensuite intéressant, au niveau de l'interprétation des résultats, d'illustrer la représentation obtenue par des informations supplémen-

taires. Sur la figure 2, les trois classes d'âge s'ordonnent comme les diplômes le long de l'axe horizontal: aux âges croissant correspondent des niveaux de diplômes décroissant. Il s'agit d'un trait structurel de la population étudiée: les moins âgés sont les plus instruits, mais ceci complique l'interprétation en terme de causalité... est-ce-que l'effet du niveau d'instruction sur le contenu et la forme des réponses libres n'est pas un effet plus direct lié à l'âge? Pour aller plus avant dans l'interprétation, il faudrait croiser les variables Âge et Diplôme, pour obtenir une partition des répondants en $3 \times 5 = 15$ colonnes... mais on sort là de l'exemple d'école qui nous sert à illustrer le fonctionnement de la méthode.

Tableau 8: Paramètres issus de l'AFC pour les éléments illustratifs

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0
VIVRE	0.75	0.57	-0.36	-0.43	0.39	0.32	0.00	0.0	0.0	0.0	0.0	0.0	0.23	0.32	0.27	0.18	0.00
MANGE	1.02	0.25	-0.47	-0.04	0.10	-0.15	0.00	0.0	0.0	0.0	0.0	0.0	0.86	0.01	0.04	0.09	0.00
CONFORT	0.75	0.56	-0.28	-0.66	-0.01	0.17	0.00	0.0	0.0	0.0	0.0	0.0	0.14	0.89	0.00	0.05	0.00
RESSENTENTE	1.03	0.15	-0.07	-0.04	0.27	-0.24	0.00	0.0	0.0	0.0	0.0	0.0	0.04	0.01	0.50	0.45	0.00
FREQUENCES ILLUSTRATIVES																	
AUT - AUTRES DIVERS	1.07	2.95	-1.03	-0.23	0.49	0.57	0.00	0.0	0.0	0.0	0.0	0.0	0.36	0.02	0.08	0.11	0.00
-30 - -30 MOINS DE 30 ANS	26.56	0.13	-0.14	0.02	-0.22	0.05	0.00	0.0	0.0	0.0	0.0	0.0	0.14	0.00	0.36	0.02	0.00
-50 - -50 DE 30 A 50 ANS	35.28	0.05	-0.02	-0.07	-0.06	0.01	0.00	0.0	0.0	0.0	0.0	0.0	0.01	0.12	0.14	0.00	0.00
+50 - +50 PLUS DE 50 ANS	38.44	0.14	0.23	0.02	0.04	0.00	0.00	0.0	0.0	0.0	0.0	0.0	0.37	0.00	0.01	0.00	0.00

3 Analyse des Correspondances Multiples

L'analyse des correspondances multiples permet de décrire de vastes tableaux binaires, dont les fichiers d'enquêtes socio-économiques constituent un exemple privilégié: les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs milliers) ; les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions. On peut faire remonter les principes de cette méthode à [GUTTMAN 1941], mais aussi à [BURT 1950] ou à [HAYASHI 1956].

Il s'agit en fait d'une simple extension du domaine d'application de l'analyse des correspondances, avec cependant des procédures de calcul et des règles d'interprétation spécifiques.

Cette extension se fonde sur la propriété suivante: On pose à k individus deux questions ayant respectivement n et p modalités de réponses, et l'on suppose que pour chaque question, les différentes modalités de réponses s'excluent mutuellement (une seule réponse possible).

Il est alors équivalent de soumettre à l'analyse des correspondances la table de contingence (n,p) croisant les deux questions, ou d'analyser le tableau binaire Z à k lignes et (n+p) colonnes qui décrit les réponses (chaque ligne de Z comporte deux "1" dans les colonnes correspondant aux réponses choisies, et n+p-2 "0").

L'analyse de Z est plus coûteuse, mais plus intéressante, car elle se généralise immédiatement au cas de plus de deux questions.

Le tableau 9 représente ainsi le cas de trois questions ayant respectivement 4, 3, 4 modalités de réponse. Le tableau "disjonctif complet" **Z** comporte donc trois blocs. Les réponses des 12 individus peuvent alors être codifiées dans un tableau réduit **R**, qui contient simplement les numéros des modalités. Les procédures de calcul n'utiliseront en fait que ce tableau peu encombrant.

Tableau 9: Tableaux **R** (Codage réduit), **Z** (disjonctif complet) et **B**=**Z**'**Z** (de Burt).

$$\mathbf{R} = \begin{pmatrix} 2 & 3 & 4 \\ 2 & 1 & 3 \\ 3 & 1 & 2 \\ 4 & 2 & 4 \\ 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 1 & 1 \\ 1 & 1 & 1 \\ 4 & 1 & 2 \\ 2 & 2 & 3 \\ 3 & 2 & 2 \\ 4 & 1 & 4 \end{pmatrix} \quad \mathbf{Z} = \begin{pmatrix} 0&1&0&0 & 0&0&1 & 0&0&0&1 \\ 0&1&0&0 & 1&0&0 & 0&0&1&0 \\ 0&0&1&0 & 1&0&0 & 0&1&0&0 \\ 0&0&0&1 & 0&1&0 & 0&0&0&1 \\ 1&0&0&0 & 0&1&0 & 0&0&1&0 \\ 0&1&0&0 & 0&1&0 & 0&0&1&0 \\ 0&0&1&0 & 1&0&0 & 1&0&0&0 \\ 1&0&0&0 & 1&0&0 & 1&0&0&0 \\ 0&0&0&1 & 1&0&0 & 0&1&0&0 \\ 0&1&0&0 & 0&1&0 & 0&0&1&0 \\ 0&0&1&0 & 0&1&0 & 0&1&0&0 \\ 0&0&0&1 & 1&0&0 & 0&0&0&1 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 2&0&0&0 & 1&1&0 & 1&1&1&1 \\ 0&4&0&0 & 1&1&1 & 0&0&3&1 \\ 0&0&3&0 & 2&1&0 & 1&2&0&0 \\ 0&0&0&3 & 2&1&0 & 0&1&0&2 \\ & & & & & & & & & & \\ 1&1&2&2 & 6&0&0 & 2&2&1&1 \\ 1&1&1&1 & 0&5&0 & 0&1&3&2 \\ 0&1&0&0 & 0&0&1 & 0&0&0&1 \\ & & & & & & & & & & \\ 1&0&1&0 & 2&0&0 & 2&0&0&0 \\ 1&0&2&1 & 2&1&0 & 0&3&0&0 \\ 1&3&0&0 & 1&3&0 & 0&0&4&0 \\ 1&1&0&2 & 1&2&1 & 0&0&0&4 \end{pmatrix}$$

Toujours sur le tableau 9, le troisième tableau **B** est le produit du tableau disjonctif **Z** par son transposé:

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z}$$

Ce dernier tableau est symétrique (on l'appelle **tableau de Burt** ou tableau de correspondance multiple). Il contient 9 blocs: les blocs diagonaux sont des matrices diagonales, dont les éléments diagonaux sont les effectifs de réponses correspondant à chaque modalité. Les trois blocs distincts parmi les blocs restants ne sont autres que les trois tables de contingence croisant les trois questions deux à deux.

Comme nous l'avons annoncé plus haut, l'analyse des correspondances du tableau **Z** donne les mêmes axes factoriels normés que celle du tableau **B**, c'est-à-dire finalement les mêmes graphiques de proximités, aux échelles près; mais les valeurs propres homologues sont distinctes: à la valeur propre l de l'analyse de **Z** correspond la valeur propre λ^2 de l'analyse de **Z**'**Z**.

Dans le cas de deux questions, le tableau **R** n'a que deux colonnes, le tableau **Z** est formé de deux blocs, et **B** en comporte quatre.

Les deux analyses précédentes donnent également la même représentation que l'analyse des correspondances de la petite table de contingence croisant les deux questions (bloc non-diagonal de **B**), avec, cette fois, la valeur propre $(2\lambda - 1)^2$.

Tableau 10: Structure de bases attitudes et opinions
Liste des questions actives et illustratives

```
                6 QUESTIONS ACTIVES
                24 MODALITES CORRESPONDANTES
|-------------------------------------------------------------------------------------------------------------|
SEXE DE L'ENQUETE(E)    ( 2 MODALITES )
|  SEXM - MASCULIN          |  SEXF - FEMININ
|-------------------------------------------------------------------------------------------------------------|
STATUT MATRIMONIAL      ( 5 MODALITES )
|  CELI - CELIBATAIRE       |  MARI - MARIE(E)          |  CONC - CONCUBINAGE       |  DIV - SEPARE (E) DIV
|  VEUF - VEUF              |                           |                           |
|-------------------------------------------------------------------------------------------------------------|
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE    ( 5 MODALITES )
|  SDIP - AUCUN             |  CEP - CEP                |  BEPC - BEPC-BEPS-BE      |  BACC - BACCALAUREAT
|  UNIV - UNIVERSITE        |                           |                           |
|-------------------------------------------------------------------------------------------------------------|
INDICATEUR D'EQUIPEMENT EN 3 CLASSES            ( 3 MODALITES )
|  EQ01 - 0 - 2             |  EQ02 - 3 - 4             |  EQ03 - 5-6-7-8           |
|-------------------------------------------------------------------------------------------------------------|
CATEGORIE SOCIO-PROFESSIONNELLE EN 6 CLASSES    ( 6 MODALITES )
|  CS01 - CAD.SUP.MOY. LIB. |  CS02 - EMPLOYE           |  CS03 - OUVRIER           |  CS04 - MENAGERE
|  CS05 - RETRAITE          |  CS06 - AUTRE             |                           |
|-------------------------------------------------------------------------------------------------------------|
AGE DE L'ENQUETE(E)                             ( 3 MODALITES )
|  AGE1 - MOINS DE 29 ANS   |  AGE2 - 30 - 49 ANS       |  AGE3 - 50 ANS ET PLUS    |
|-------------------------------------------------------------------------------------------------------------|

                5 VARIABLES ILLUSTRATIVES ( 16 MODALITES)

|-------------------------------------------------------------------------------------------------------------|
LA FAMILLE EST LE SEUL ENDROIT OU L'ON SE SENT BIEN?    ( 3 MODALITES )
|  FAM1 - OUI               |  FAM2 - NON               |  FAM3 - NSP               |
|-------------------------------------------------------------------------------------------------------------|
OPINION SUR LE MARIAGE                          ( 3 MODALITES )
|  MAR1 - INDISSOLUBLE      |  MAR2 - DISSOUT PB.GRAVE  |  MAR3 - DISSOUT SI ACCORD |
|-------------------------------------------------------------------------------------------------------------|
AVEZ-VOUS L'USAGE D'UN ORDINATEUR DOMESTIQUE?   ( 2 MODALITES )
|  ORD1 - OUI               |  ORD2 - NON               |                           |
|-------------------------------------------------------------------------------------------------------------|
REGARDEZ-VOUS LA TELEVISION?                    ( 4 MODALITES )
|  TEL1 - TOUS LES JOURS    |  TEL2 - ASSEZ SOUVENT     |  TEL3 - PAS TRES SOUVENT  |  TEL4 - JAMAIS
|-------------------------------------------------------------------------------------------------------------|
LES DECOUVERTES SCIENTIFIQUES AMELIORENT-ELLES LA VIE?   ( 4 MODALITES )
|  SCI1 - UN PEU            |  SCI2 - BEAUCOUP          |  SCI3 - PAS DU TOUT       |  SCI4 - NE SAIT PAS
|-------------------------------------------------------------------------------------------------------------|
```

On notera que dans ce cas les lignes et les colonnes de la table de contingence ne sont que des colonnes pour la matrice **Z**: la représentation simultanée des lignes et des colonnes n'est donc pas un simple artifice graphique: elle trouve une justification naturelle avec cette présentation de l'analyse des correspondances.

Un exemple: *Structure de base de l'échantillon de "l'enquête sur les conditions de vie et aspirations des Français en 1984".*

Cet exemple illustre une utilisation assez courante de l'analyse des correspondances multiples pour le traitement des données d'enquêtes. L'exemple précédent qui concernait une réponse

libre de la même enquête statistique a attiré l'attention sur le fait que derrière un effet "niveau d'éducation" pouvait se cacher un effet d'âge... (les moins instruits sont en moyenne les plus âgés). Etant donnée la structure actuelle de la population française, ces mêmes effets, on le sait, ne sont pas indépendants du sexe, du niveau de vie... d'où l'idée de décrire le réseau d'interrelations entre toutes ces caractéristiques de base, puis de positionner les autres thèmes de l'enquête en tant qu'éléments illustratifs.

Toujours pour faciliter la présentation des résultats, l'échelle réelle de l'application sera réduite: Le tableau Z aura cependant 2000 lignes et 24 colonnes actives représentant les modalités de réponses à 6 questions (tableau 10). Il comprendra aussi 5 questions illustratives, décrivant quelques attitudes et opinions, et également quelques modalités illustratives complémentaires constituées par les formes graphiques de l'exemple précédent occupant des positions significatives (l'élément z_{ij} correspondant de Z vaut 1 si l'individu i a utilisé le mot j pour la réponse libre précitée, et 0 sinon.

Tableau 11: Croisements deux à deux des questions actives

TABLEAU DE BURT

	SEXH	SEXF	CELI	MARI	CONC	DIVO	VEUF	SDIP	CEP	BEPC	BACC	UNIV	EQ01	EQ02	EQ03	CS01	CS02	CS03	CS04	CS05	CS06	AGE1	AGE2	AGE3
SEXH	963.																							
SEXF	0.1037																							
CELI	208.	140.	348.																					
MARI	608.	583.	0.1191.																					
CONC	63.	64.	0.	0.	127.																			
DIVO	54.	65.	0.	0.	0.	119.																		
VEUF	30.	185.	0.	0.	0.	0.	215.																	
SDIP	313.	224.	55.	242.	24.	21.	94.	436.																
CEP	331.	313.	44.	469.	22.	31.	78.	0.	644.															
BEPC	175.	221.	91.	223.	32.	26.	24.	0.	0.	396.														
BACC	143.	198.	114.	154.	32.	25.	16.	0.	0.	0.	361.													
UNIV	102.	81.	44.	103.	17.	14.	3.	0.	0.	0.	0.	183.												
EQ01	291.	297.	148.	235.	67.	58.	60.	168.	147.	123.	102.	49.	386.											
EQ02	530.	601.	136.	757.	56.	54.	128.	236.	439.	219.	164.	75.	0.	1131.										
EQ03	142.	139.	46.	199.	4.	7.	27.	36.	58.	56.	75.	60.	0.	0.	281.									
CS01	184.	118.	73.	170.	28.	27.	4.	16.	32.	56.	99.	105.	75.	165.	62.	302.								
CS02	66.	171.	63.	135.	29.	31.	10.	25.	64.	103.	55.	10.	87.	151.	19.	0.	259.							
CS03	281.	56.	71.	206.	70.	39.	2.	107.	161.	53.	15.	1.	137.	109.	11.	0.	0.	337.						
CS04	10.	376.	4.	292.	13.	12.	64.	95.	147.	65.	55.	22.	86.	238.	62.	0.	0.	0.	384.					
CS05	212.	160.	17.	222.	4.	15.	116.	130.	163.	46.	34.	19.	76.	221.	75.	0.	0.	0.	0.	372.				
CS06	190.	158.	121.	166.	25.	17.	19.	69.	97.	73.	83.	36.	129.	167.	52.	0.	0.	0.	0.	0.	369.			
AGE1	233.	276.	227.	196.	69.	17.	0.	77.	85.	154.	145.	48.	232.	242.	35.	69.	106.	114.	69.	0.	151.	509.		
AGE2	360.	345.	76.	497.	49.	65.	19.	107.	247.	140.	122.	89.	190.	420.	95.	181.	113.	159.	130.	0.	115.	0.	705.	
AGE3	370.	416.	43.	496.	5.	39.	197.	252.	312.	102.	74.	46.	166.	469.	151.	52.	38.	63.	177.	372.	82.	0.	0.	796.
	SEXH	SEXF	CELI	MARI	CONC	DIVO	VEUF	SDIP	CEP	BEPC	BACC	UNIV	EQ01	EQ02	EQ03	CS01	CS02	CS03	CS04	CS05	CS06	AGE1	AGE2	AGE3

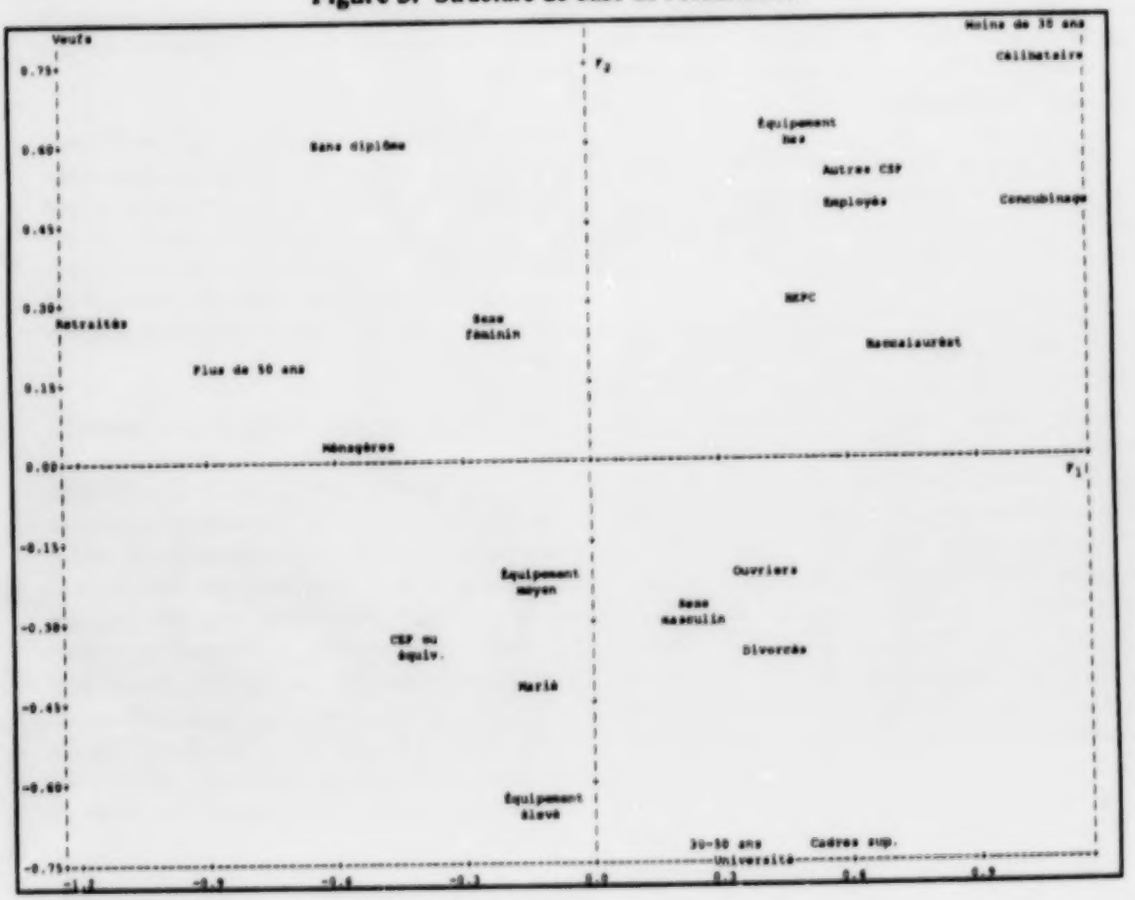
L'enquête présentée ici est annuelle depuis 1978, mais les 2000 individus interrogés ici (Français métropolitains de plus de 18 ans) l'ont été en 1984. Le questionnaire très vaste traite notamment de la famille, de la santé, du travail, des transports, de l'énergie. Dans une exploita-

tion en vrai grandeur, il n'est pas rare de positionner en éléments illustratifs plusieurs centaines de modalités.

Lecture du tableau 11

Le tableau 11 ne représente que la moitié inférieure du tableau de Burt (qui est, rappelons-le, symétrique) relatif aux 6 questions actives . On trouve dans ce tableau tous les tableaux de contingence (il y en a 15) croisant les 6 questions actives deux à deux. Sur la diagonale se trouvent les questions croisées avec elles-mêmes, et donc les effectifs correspondant à chaque modalité: on peut ainsi lire qu'il y a 963 hommes et 1037 femmes dans l'échantillon. Sur 215 veufs, il y a 30 hommes et 185 femmes.

Figure 3: Structure de base de l'échantillon



Lecture de la Figure 3

La structure du "nuage" des modalités actives est décrite par le plan factoriel de la Figure 3, qui résume donc 15 tableaux croisés. Malgré le caractère sommaire des variables utilisées, les liaisons entre niveau de diplôme, âge, sexe et équi-

pement sont plus complexes que l'on ne l'imagine a priori.

Un effet de cycle de vie, et aussi de génération se cache derrière des variables aussi élémentaires que le sexe: La catégorie "sexe féminin" est lestée par un surnombre de personnes âgées sans diplôme. Il faut donc en tenir compte chaque fois que l'on interprète un éventuel "effet" du sexe sur telle opinion ou attitude. L'équipement des ménages augmente pendant une partie du cycle de vie, mais augmente aussi avec le niveau de diplôme; or les jeunes sont plus diplômés... il y a donc conflit entre deux effets: certaines catégories peu diplômées sont apparemment mieux équipées simplement parce qu'elles sont plus âgées en moyenne.

La figure 3 n'est cependant qu'une approximation plane d'un ensemble de proximités qui ne peuvent pas toujours être représentées dans un plan... il y a donc certaines déformations.

Prenons un exemple: le point "divorcés" est plus proche du point "sexe masculin" que du point "sexe féminin", alors qu'un rapide calcul de pourcentages fait à partir du tableau 11 nous montre qu'il y a une proportion légèrement plus forte de femmes divorcées. Mais les personnes divorcées appartiennent massivement à la classe d'âge intermédiaire, sont plus diplômées que la moyenne, sont surreprésentées chez les cadres supérieurs et les ouvriers (professions masculines)... en somme la position du point "divorcés" résulte d'un "compromis", mais c'est un "compromis" instructif: les femmes divorcées ont surtout des caractéristiques de la population masculine.

Même si elle peut donner lieu à quelques commentaires, la figure 3 n'est pas un résultat en soi: elle n'est qu'une préparation de l'information destinée à accueillir les variables illustratives. Auparavant, disons quelques mots de la validité de la représentation.

Validité de la représentation

On évitera ici d'utiliser les pourcentages de variance pour caractériser les axes: dans le cas des correspondances multiples, ils n'ont pas le même sens que lorsqu'il s'agit d'une table de contingence: le codage binaire introduit un "bruit" qui réduit la part d'explication attachée à chaque valeur propre.

Ici, les deux premiers pourcentages sont respectivement de 13% et 9%, pour des valeurs propres valant 0,39 et 0,27. En pratique, la stabilité du plan factoriel s'éprouve par des techniques de simulation (perturbations aléatoires du tableau de données) ou de validation par échantillon-test.

Ici, la stabilité du "pattern" (cet anglicisme permettant de désigner rapidement quelque chose d'intermédiaire entre une configuration et une structure) est acquise, puisque l'enquête est annuelle, et qu'il a été possible de vérifier la permanence de la représentation observée par l'analyse d'échantillons indépendants correspondant à des années différentes.

Positionnement des variables illustratives

La représentation simultanée des lignes et des colonnes liée à l'analyse des correspondances n'a pas été utilisée: seules les colonnes du tableau Z ont été portées sur la figure 3. Les 2000 points-lignes n'ont pas été tracés, pour des raisons d'encombrement graphique, mais surtout parce que ces individus sont anonymes... ils ne présentent de l'intérêt que par l'intermédiaire de leurs caractéristiques. Ce sont précisément les autres informations disponibles sur les personnes interrogées qui vont être projetées en éléments illustratifs.

Le fait que le tableau Z ne comporte que des "0" et des "1" va permettre une interprétation particulièrement simple de la disposition relative des colonnes illustratives:

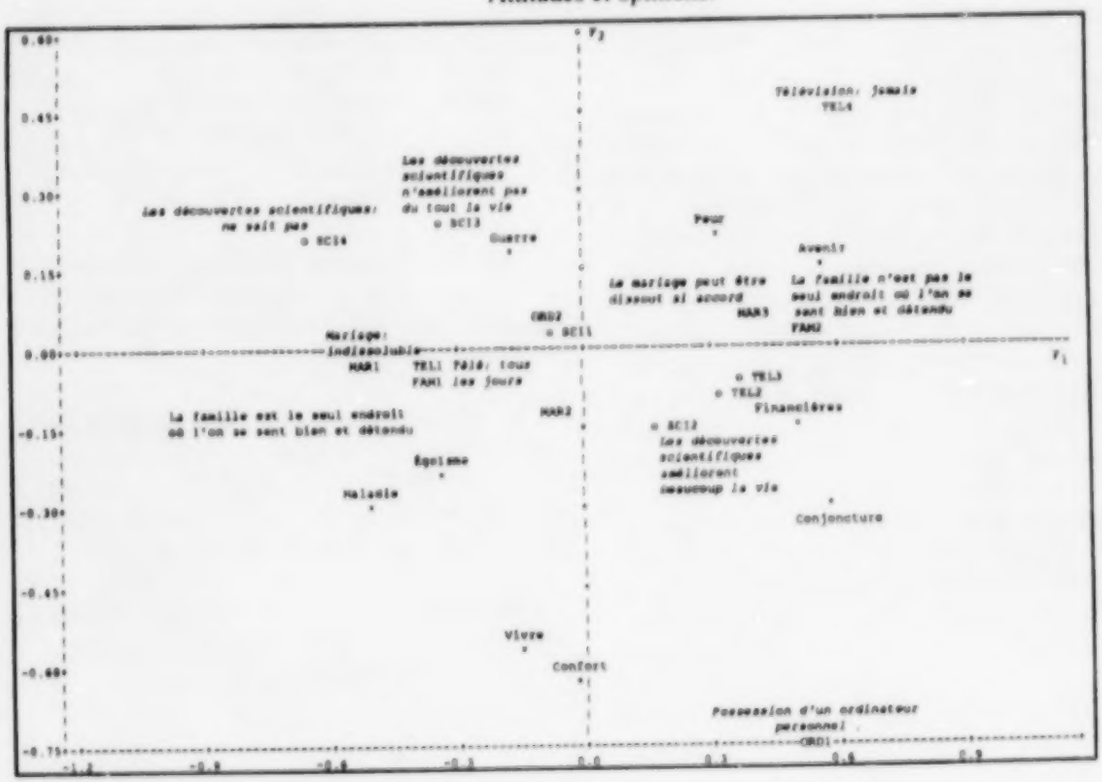
La formule de transition (2) précédente nous montre que la coordonnée F_j d'une colonne j (réponse illustrative) s'obtient en multipliant par β la moyenne arithmétique des coordonnées des individus qui ont choisi la réponse j.

Lecture de la figure 4

On lit tout d'abord sur la figure 4 le positionnement des réponses aux questions dont les libellés complets figurent dans le tableau 10. Ces réponses sont donc projetées comme éléments illustratifs sur le plan factoriel de la figure 3 (les modalités des variables actives n'ont pas été reproduites sur la figure 4 afin d'en alléger la lecture).

Les caractéristiques des personnes qui répondent sont visibles immédiatement dans un cadre qui tient compte des interrelations existant entre ces caractéristiques. Les consultations classiques de tableaux croisés sont en effet hypothéquées par le fait "qu'une variable peut en cacher une autre"... elles sont de plus largement redondantes lorsque les caractéristiques successives sont liées entre elles. Le système de projection de variables supplémentaires permet donc d'économiser du temps et d'éviter des erreurs d'interprétation.

Figure 4: Structure de base modalités illustratives.
Attitudes et opinions.



La confrontation des figures 3 et 4 permet d'énoncer les résultats suivants:

Alors que les opinions sur la famille et le mariage se développent le long de l'axe horizontal lié à l'âge et au diplôme (les jeunes instruits ayant des opinions "modernistes" par opposition aux personnes plus âgées et peu instruites, plus "traditionalistes"), les opinions sur l'effet des découvertes scientifiques se positionnent le long d'une diagonale qui fait plus intervenir le sexe et le niveau de vie, les plus chauds partisans de la science appartenant à la classe d'âge moyenne.

L'écoute de la télévision est nettement liée à l'âge, et corrélativement au niveau d'instruction, alors que la possession d'un ordinateur personnel caractérise les cadres supérieurs et les diplômés d'université d'âge moyen.

La seconde série de variables illustratives portée sur la figure 3 est constituée par 9 formes graphiques utilisées par les mêmes individus dans leurs réponses à la question ouverte de l'exemple 1 ("quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant?"). Ces neuf formes ont été sélectionnées en raison de leurs positions excentrées, et donc caractéristiques. Lors d'une application en vraie grandeur, ce sont des centaines de mots que l'on projeterait de cette façon. On retrouve des termes tels que "financières", "conjoncture" dans une zone où les personnes interrogées sont plutôt jeunes et instruites, "vivre" et "confort" dans une zone d'âge moyen à niveau de vie élevé, "avenir" du côté des plus jeunes, une proximité qui peut être intéressante entre "guerre" et une attitude négative vis-à-vis de la science. Toutes ces proximités suggèrent en fait des conjectures à vérifier ensuite sur les données initiales.

On peut résumer la démarche suivie lors de cette application de l'analyse des correspondances multiples par les deux étapes:

1) Établissement d'une typologie bidimensionnelle de la population selon un point de vue caractérisé par un ensemble relativement homogène de variables actives (ici: les caractéristiques de base).

2) Identification des sous-groupes de cette population à partir de toutes les autres informations pertinentes disponibles. En fait, c'est l'ordinateur lui même qui sélectionnera les variables supplémentaires ayant des coordonnées significatives sur les axes factoriels, ce qui permet d'envisager des explorations systématiques, avec de nombreux croisements de variables.

4 Complémentarité de la Classification

Dans le cas du traitement statistique des fichiers d'enquêtes en vraie grandeur, la démarche précédente fondée sur des représentations graphiques a au moins deux défauts:

1) Les visualisations sont limitées à deux, ou en général, à très peu de dimensions, alors que le nombre d'axes significatifs[1] peut souvent atteindre 8 ou 10.
2) Ces visualisations peuvent inclure des centaines de points, et donner lieu à des graphiques chargés ou illisibles.

Il faut donc à ce stade faire appel de nouveau aux capacités de gestion et de calcul de l'ordinateur pour compléter, alléger et clarifier la présentation des résultats.

L'utilisation conjointe de la classification automatique et des analyses précédentes permet de remédier à ces lacunes. Lorsqu'il y a trop de points sur un graphique, il paraît utile de procéder à des regroupements en familles homogènes.

Mais les algorithmes utilisés pour ces regroupements fonctionnent de la même façon, que les points soient situés dans un espace à deux ou à dix dimensions.

---

1 *On appelle ici axe significatif, un axe qui explique une part de dispersion ne pouvant être imputée au hasard. On peut déterminer le nombre d'axes significatifs en utilisant une procédure de simulation.*

Autrement dit, l'opération va présenter un double intérêt: **allégement des sorties graphiques** d'une part, **prise en compte de la dimension réelle du nuage de points** d'autre part.

Une fois les individus regroupés en classes, il est facile d'obtenir une description automatique de ces classes: on peut en effet, pour les variables numériques comme pour les variables nominales, calculer des statistiques d'écarts entre les valeurs internes à la classe et les valeurs globales; on peut également convertir ces statistiques en **valeurs-test** et opérer un tri sur ces valeurs-test On obtient finalement, pour chaque classe, les modalités et les variables les plus caractéristiques.

Après quelques mots sur les algorithmes de classification, on présentera un exemple d'application qui prolongera l'analyse des correspondances multiples précédente, et qui illustrera cette nouvelle présentation compacte de l'information.

4.1 Les algorithmes de classification

L'algorithme de classification qui nous paraît le plus adapté au partitionnement d'un ensemble comprenant des milliers d'individu est un *algorithme mixte* procédant en trois phases:

1) **Partitionnement initial** en quelques dizaines de classes (par une technique du type "nuées dynamiques" ou "k-means"). On peut résumer sans trop les trahir ces techniques en quelques phrases.

   On commence par tirer au hasard des individus qui seront des centres provisoires de classes. Puis, on affecte tous les individus au centre provisoire le plus proche. (proche au sens d'une distance telle que la distance du Khi-2 définie au paragraphe 2). On construit ainsi une partition de l'ensemble des individus. On calcule de nouveau des centres provisoires, qui sont maintenant des "centres" (points moyens par exemple) des classes qui viennent d'être obtenues, et on itère le processus, autrement dit on affecte de nouveau tous les individus à ces centres, ce qui induit une nouvelle partition, etc... le processus se stabilise nécessairement, mais la partition obtenue dépendra en général du choix initial des centres.

2) **Agrégation hiérarchique des classes obtenues**

   Ces techniques sont assez coûteuses si elles s'appliquent à des milliers d'individus, c'est pourquoi il faut réduire la dimension du problème en opérant un regroupement préalable en quelques dizaines de classes.

   Le principe de l'agrégation hiérarchique est également simple: partant de n individus, on agrège les deux individus les plus proches, et l'on considère le couple agrégé comme un nouvel indi-vidu (ce qui peut être fait de différentes façons, d'où un nombre impressionnant de "variantes" de cette technique); on est ramené au problème précé-dent avec maintenant (n-1) individus.. on agrège à nouveau les deux plus proches, ce qui réduit à (n-2) le nombre d'individus, et l'on itère le processus jusqu'à n'avoir plus qu'un seul individu, qui sera le sommet de "l'arbre hiérarchique". L'intérêt de cet arbre (ou dendrogramme) est qu'il peut donner une idée du nombre de classes existant effectivement dans la population. Chaque coupure d'un tel arbre va donner une partition, ayant d'autant moins de classes que l'on coupe près du sommet.

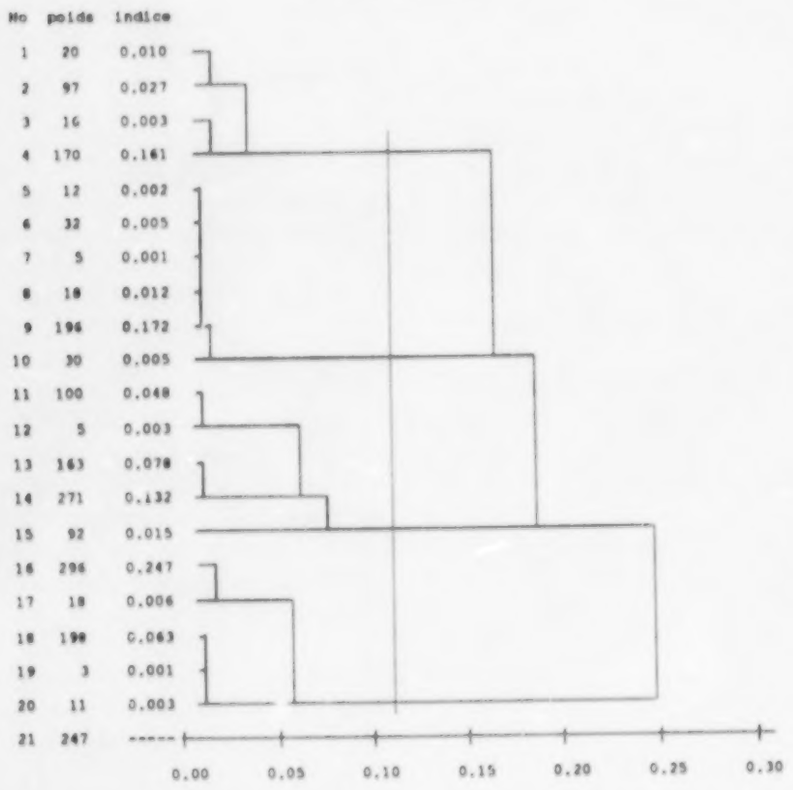
3) **Coupure de l'arbre** (en général après une inspection visuelle) et optimisation de la partition obtenue par réaffectations.

Plus on agrège de points, autrement dit plus on se rapproche du sommet de l'arbre, plus la distance entre les deux classes les plus proches est grande. On peut "indexer" les "noeuds" de l'arbre par cette plus petite distance, comme le montre la figure 5, de façon à ce que de grandes branches correspondent à

des augmentations importantes de cette distance. En coupant l'arbre au niveau d'un saut important de cet indice, on peut espérer obtenir une partition de bonne qualité, car les individus regroupés auparavant étaient proches, et ceux regroupés après la coupure seront nécessairement éloignés, ce qui est la définition d'une bonne partition...

Mais une telle partition, en k classes par exemple, n'est pas la meilleure possible, car l'algorithme de classification hiérarchique n'a malheureusement pas la propriété de donner à chaque étape une partition optimale. On peut encore améliorer la partition obtenue par réaffectation des individus comme indiqué au paragraphe 4.1 ci-dessus. Malgré la relative complexité de la procédure, on ne peut être assuré d'avoir trouvé la "meilleure partition en k classes".

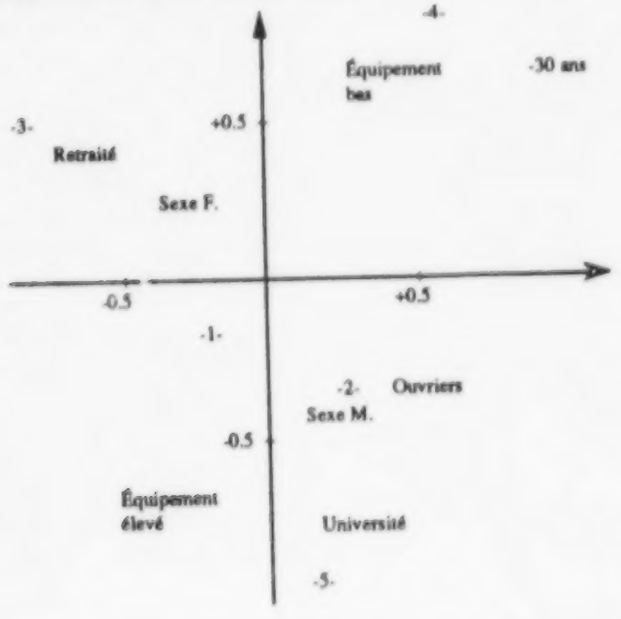
Figure 5: Dendrogramme.
Coupure en classes



Guide de lecture de la figure 5

–La première colonne contient les numéros de 21 classes obtenues par un partitionnement direct.
–La seconde colonne donne les effectifs de ces classes.
–La troisième colonne donne la valeur de l'indice pour laquelle cette classe est agrégée à une autre classe.
–Les valeurs approchées de cet indice figurent également en abscisse: elles servent à repérer les niveaux d'agrégation.

Figure 6: Positions des 5 centres de classes sur la figure 3



Les analyses complètes se dérouleront en définitive de la façon suivante:

1) Choix des éléments actifs (i.e: choix d'un point de vue. On peut décrire les individus du point de vue de leurs caractéristiques de base, mais aussi à partir d'un thème particulier: habitudes de consom-mation, opinions politiques, etc...).
2) Analyse des correspondances simples ou multiples.
3) Positionnement des éléments illustratifs. On projettera ainsi toute l'information disponible

de nature à comprendre ou à interpréter la typologie induite par les éléments actifs.

4) Examen des premiers graphiques de plans factoriels en général limités aux points occupant les positions les plus significatives.
5) Partition de l'ensemble des individus ou observations.
6) Position sur les graphiques précédents des centres des principales classes (une partition définit toujours une variable nominale particulière, dont les modalités sont les réponses à la question: "à quelle classe appartenez-vous?"). Ces modalités peuvent donc elles aussi être projetées en éléments illustratifs sur les plans factoriels.
7) Description des classes par les modalités et les variables les plus caractéristiques.

4.2 Exemple d'application: Situations-types ou Noyaux factuels

Cet exemple est le prolongement de l'analyse des correspondances multiples précédente. Les 2000 individus enquêtés sont décrits par le même ensemble de variables actives (Tableau 10). On veut maintenant obtenir un petit nombre de groupes d'individus les plus homogènes possible vis-à-vis de leurs caractéristiques de base.

L'idée en est brièvement la suivante: on aimerait croiser des caractéristiques tels que l'âge, le sexe, la profession, le niveau d'instruction, de façon à étudier des groupes d'individus tout à fait comparables entre eux du point de vue de leur situation objective. Mais de tels croisements conduisent vite à des milliers de modalités, dont on ne sait que faire lorsque l'on étudie un échantillon de l'ordre de 1000 ou 2000 individus. De plus, les croisements ne tiennent pas compte du réseau d'interrelations existant entre ces caractéristiques: certaines sont évidentes a priori (il n'y a pas de "moins de 30 ans" retraités, d'hommes "femmes au foyer"), d'autres sont également connues a priori, mais peuvent souffrir des exceptions (il y a peu d'étudiants veufs, d'ouvriers diplômés d'université), d'autres enfin ont un caractère plus statistique (il y a plus de femmes dans les catégories employés et veufs).

En pratique, on peut espérer trouver des regroupements opératoires en une vingtaine de classes pour un échantillon de l'ordre de 2000 individus. On verra qu'un regroupement plus sommaire en cinq grandes classes n'est pas totalement dépourvu d'intérêt, en restant compatible avec le volume restreint d'un exemple pédagogique. De telles classes sont appelés situations-types ou encore noyaux factuels.

La figure 5 nous montre le "dendrogramme" obtenu en agrégeant de proche en proche les 21 groupes obtenus par le partitionnement initial décrit en a) ci-dessus. Le passage de 6 à 5 classes correspond à un saut important de l'indice (de 0.078 à 0.132), ce qui montre qu'il existe une "bonne" partition en 5 classes. Rappelons que celle-ci sera ensuite améliorée par réaffectation des individus.

La figure 6 représente à nouveau le plan factoriel, présenté sur la figure 3 jalonné par le rappel de la position approximative de certains points, et enrichi par le positionnement des centres des 5 classes qui viennent d'être calculées.

Déjà, la figure 6, confrontée avec la figure 3 plus complète, permet d'avoir une idée de la composition des 5 classes (La classe 4 est formée de jeunes, la 3 de personnes âgées, la 5 de personnes diplômées...). Mais il s'agit d'une appréciation visuelle et qualitative, hypothéquée par le fait que l'on a ici affaire à une projection plane de points qui sont dans un espace de dimension élevée.

Le tableau 12 va en revanche décrire les classes de façon précise, en comparant les pourcentages de réponses internes aux classes aux pourcentages globaux, puis en sélectionnant les modalités les plus caractéristiques (cf. encadré de lecture du tableau 12). On note que parmi les modalités les plus caractéristiques d'une classe figurent également des modalités illustratives, qui

n'ont pas participé à la fabrication des classes. Tels sont les cas des réponses sur la famille, "traditionalistes" surtout pour la classe 3, modernistes pour les classes 4 et 5; de l'écoute de la télévision, surtout élevée dans la classe 3 ; des découvertes scientifiques, plus appréciées dans la classe 5, classe également caractérisée par un taux de possession d'ordinateur domestique trois fois plus élevé qu'en moyenne.

On voit assez bien les avantages que présentent les partitions pour décrire des ensembles multidimensionnels:

La notion de classe est intuitive (groupes d'individus les plus semblables possibles). La description des classes fait appel à des classements de libellés complets et donc facile à lire, ces classements étant fondés sur des simples comparaisons de pourcentages. Il est donc plus facile de décrire des classes qu'un espace continu.

Mais l'analyse des correspondances permet de visualiser les positions relatives des classes dans l'espace, et aussi de mettre en évidence certaines variations continues ou dérivées dans cet espace qui auraient pu être masquées par la discontinuité des classes.

Les deux techniques sont donc complémentaires, et se valident mutuellement.

Guide de lecture du Tableau 12

Pour chacune des classes, les modalités les plus caractéristiques sont rangées suivant les valeurs décroissantes d'un CRITERE, qui constituent la première colonne numérique du tableau. Ce critère est, brièvement, l'analogue de la valeur absolue d'une variable normale centrée réduite, qui est significatif au seuil 5% s'il dépasse la valeur 2, ce qui est le cas pour toutes les modalités retenues ici.

La seconde colonne, POIDS, donne les effectifs des classes et des modalités.

La troisième colonne, POURCENTAGE GLOBAL, donne le pourcentage de chaque modalité dans la population globale qui s'obtient simplement en divisant la colonne poids par 2000, effectif global de l'échantillon.

La quatrième et dernière colonne donne le pourcentage INTERNE de chaque modalité. Il est nécessairement plus élevé que le pourcentage global, puisque les modalités sélectionnées sont caractéristiques des classes. C'est précisément la différence entre le pourcentage interne et le pourcentage global qui est à la base du calcul du critère.

Ainsi, la classe 2 contient 439 individus, soit 22% de la population. La modalité "sexe masculin", qui concerne 48% de l'échantillon, concerne près de 87% des personnes de cette classe. Elle vient en second rang pour caractériser la classe, après la modalité "ouvrier" de la variable "Catégorie socio-professionnelle".

Tableau 12: Description automatique des 5 noyaux factuels

DESCRIPTION DES 5 CLASSES		CRITERE	POIDS	POURCENTAGE GLOBAL	POURCENTAGE INTERNE
CLASSE1			478	23.9	
SEXE DE L'ENQUETE(E)	=FEMININ	24.8	1037	51.8	96.0
CATEGORIE SOCIO PROFESSIONNELLE	=MENAGERE	24.1	384	19.2	63.2
STATUT MATRIMONIAL	=MARIE(E)	15.5	1191	59.5	87.7
CATEGORIE SOCIO PROFESSIONNELLE	=EMPLOYE	8.6	257	12.8	25.5
INDICATEUR D'EQUIPEMENT	=MOYEN	8.0	1131	56.5	72.0
AGE DE L'ENQUETE(E)	=30 - 49 ANS	7.5	705	35.3	49.8
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=BEPC-BEPS-BE	4.7	396	19.8	27.6
OPINION SUR LE MARIAGE	=DISSOUT SI PB.GRAVE	4.5	681	34.0	42.7
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=CEP	4.3	642	32.1	40.2
LA FAMILLE...SEUL ENDROIT OU L'ON SE SENT BIEN ?	=OUI	2.8	1216	60.8	66.1
CLASSE2			439	22.0	
CATEGORIE SOCIO PROFESSIONNELLE	=OUVRIER	28.3	337	16.8	70.8
SEXE DE L'ENQUETE(E)	=MASCULIN	19.1	963	48.1	86.8
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=CEP	10.2	642	32.1	53.1
AGE DE L'ENQUETE(E)	=30 - 49 ANS	9.7	705	35.3	55.4
STATUT MATRIMONIAL	=MARIE(E)	5.1	1191	59.5	69.9
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=AUCUN	4.9	435	21.8	30.5
INDICATEUR D'EQUIPEMENT	=BAS	4.2	588	29.4	37.6
AVEZ-VOUS L'USAGE D'UN ORDINATEUR DOMESTIQUE	=NON	3.5	1949	97.4	99.3
STATUT MATRIMONIAL	=SEPARE(E) DIVORCE(E)	3.0	119	5.6	9.1
LA FAMILLE...SEUL ENDROIT OU L'ON SE SENT BIEN ?	=OUI	2.8	1216	60.8	66.5
CLASSE3			476	23.8	
AGE DE L'ENQUETE(E)	=50 ANS ET PLUS	33.2	786	39.3	99.8
CATEGORIE SOCIO PROFESSIONNELLE	=RETRAITE	31.1	372	18.6	75.2
STATUT MATRIMONIAL	=VEUF(VE)	20.3	215	10.8	40.1
LA FAMILLE...SEUL ENDROIT OU L'ON SE SENT BIEN ?	=OUI	12.9	1216	60.8	84.5
REGARDEZ-VOUS LA TELEVISION	=TOUS LES JOURS	12.4	1140	57.0	80.5
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=AUCUN	10.3	435	21.8	40.1
OPINION SUR LE MARIAGE	=INDISSOLUBLE	10.0	466	23.3	41.4
AVEZ-VOUS L'USAGE D'UN ORDINATEUR DOMESTIQUE	=NON	5.2	1949	97.4	99.8
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=CEP	3.6	642	32.1	38.9
INDICATEUR D'EQUIPEMENT	=ELEVE	3.0	281	14.0	18.3
LES DEC. SCIENTIFIQUES AMELIORENT-ELLES LA VIE?	=PAS DU TOUT	2.6	219	10.9	14.3
CLASSE4			309	15.4	
AGE DE L'ENQUETE(E)	=MOINS DE 29 ANS	24.1	509	25.4	88.3
STATUT MATRIMONIAL	=CELIBATAIRE	21.3	348	17.4	70.6
CATEGORIE SOCIO PROFESSIONNELLE	=AUTRE	14.5	348	17.4	51.8
LA FAMILLE...SEUL ENDROIT OU L'ON SE SENT BIEN ?	=NON	12.2	783	39.1	70.9
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=BACCALAUREAT B.SUP	10.8	339	16.9	41.4
OPINION SUR LE MARIAGE	=DISSOUT SI ACCORD	10.1	779	38.9	65.4
INDICATEUR D'EQUIPEMENT	=BAS	9.6	588	29.4	53.7
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=BEPC-BEPS-BE	8.1	396	19.8	38.5
CATEGORIE SOCIO PROFESSIONNELLE	=EMPLOYE	8.1	257	12.8	29.1
STATUT MATRIMONIAL	=CONCUBINAGE	6.5	127	6.3	16.2
REGARDEZ-VOUS LA TELEVISION	=JAMAIS	5.4	108	5.4	12.9
REGARDEZ-VOUS LA TELEVISION	=PAS TRES SOUVENT	4.7	340	17.0	26.9
CLASSE5			298	14.9	
CATEGORIE SOCIO PROFESSIONNELLE	=CAD.SUP.MOY. P.LIB.	26.3	302	15.1	83.2
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=UNIVERSITE G.ECOLES	18.6	180	9.0	48.3
AGE DE L'ENQUETE(E)	=30 - 49 ANS	11.7	705	35.3	66.4
LA FAMILLE...SEUL ENDROIT OU L'ON SE SENT BIEN ?	=NON	8.2	783	39.1	61.1
INDICATEUR D'EQUIPEMENT	=ELEVE	7.1	281	14.0	28.9
SEXE DE L'ENQUETE(E)	=MASCULIN	6.8	963	48.1	66.1
LES DEC. SCIENTIFIQUES AMELIORENT-ELLES LA VIE?	=BEAUCOUP	6.6	644	32.2	49.3
DIPLOME D'ENSEIGNEMENT GENERAL LE PLUS ELEVE	=BACCALAUREAT B.SUP	5.6	339	16.9	29.2
OPINION SUR LE MARIAGE	=DISSOUT SI ACCORD	5.3	779	38.9	53.0
AVEZ-VOUS L'USAGE D'UN ORDINATEUR DOMESTIQUE	=OUI	5.0	51	2.5	7.7
REGARDEZ-VOUS LA TELEVISION	=PAS TRES SOUVENT	4.5	340	17.0	26.5
STATUT MATRIMONIAL	=SEPARE(E) DIVORCE(E)	4.3	119	5.9	12.1

Bibliographie sommaire

BENZÉCRI J-P. & coll. (1973) - *La Taxinomie*, Vol. I, *L'Analyse des Correspondances*, Vol. II, Dunod, Paris.

BOUROCHE J.M., SAPORTA G. (1980) - *L'Analyse des Données*, coll."Que sais-je", n°1854, P.U.F. Paris .

BURT C. (1950) - The factorial Analysis of Qualitative Data. *British J. of stat. psychol.* vol 3, n°3, p 166-185.

CAILLEZ F., PAGES J.-P. (1976) - Introduction à l'Analyse des Données. S.M.A.S.H., Paris.

CIBOIS P. (1984) - *Analyse des Données en Sociologie*. P.U.F. Paris.

CREDOC (1980-1988) - *Rapports Annuels de l'Enquête sur les Conditions de Vie et Aspirations des Français*, publiés sous la direction. de L. Lebart, avec la collaboration de F. Boscher, E. Brian, F. Chateau, G. Drouault, C. Duflos, B. Ezvan, M. Grignon, F. Gros, L. Haeusler, Y. Houzel, L. Laurent, P. Pleuvret.

DIDAY E. (1971) - La Méthode des Nuées Dynamiques. *Revue Stat. Appl.*,vol.19, n°2, p 19-34.

ESCOUFIER Y. (1985) - L'Analyse des Correspondances, ses Propriétés, ses Extensions. *Bull. of the Int. Stat. Inst.*, 4, 28-2.

FÉNELON J.P. (1981) - *Qu'est-ce-que l'Analyse des Données?* Lefonen, Paris.

GIFI A. (1981) - *Non Linear Multivariate Analysis*. Department of Data Theory. Faculty of Social Sciences. Univ. of Leiden, The Netherlands.

GUTTMAN L. (1941) - The Quantification of a Class of Attributes: a Theory and Method of a Scale Construction. In *"The prediction of personal adjustment"* (P. Horst, ed.) p 251 -264, SSCR New-York.

HAYASHI C. (1956) - Theory and Examples of Quantification. (II) *Proc. of the Institute of Stat. Math.* 4 (2) p 19-30.

LEBART L. (1987) - Conditions de Vie et Aspirations des Français, Évolution et Structure des Opinions de 1978 à 1986. *Futuribles*, sept 1987, p 25-56.

LEBART L., MORINEAU A., FÉNELON J.-P. (1979) - *Traitement des Données Statistiques*. Dunod, Paris.

LEBART L., MORINEAU A., WARWICK K. (1984) - *Multivariate Descriptive Statistical Analysis*. J. Wiley and sons. New York.

LEBART L., SALEM A. (1988) - *Analyse Statistique des Données Textuelles*, Dunod, Paris.

NISHISATO S. (1980) - *Analysis of Categorical Data. Dual Scaling and its Application*. Univ. of Toronto Press.

NISHISATO S. (1986) - *Quantification of Categorical Data: A Bibliography* 1975-1986. Microstats, Toronto.

SPAD.N (1987) - *Système Portable pour l'Analyse des Données*. (SPAD.T (1988) - Système Portable pour l'Analyse des Données Textuelles). CISIA Ed., 25 Ave de l'Europe 92310 Sèvres.

TENENHAUS M., YOUNG F. W. (1985) - An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and other Methods for Quantifying Categorical Multivariate Data. *Psychometrika*, vol 50; p 91-119.

VOLLE M. (1980) - *Analyse des Données*. Economica, Paris.

L'UTILISATION DE L'ANALYSE DES CORRESPONDANCES MULTIPLES DANS L'ÉVALUATION DE L'ACIDIFICATION DES EAUX LACUSTRES DU QUÉBEC

Marius Lachance
INRS-Eau

1 Introduction

Dans l'est de l'Amérique du Nord, le transport atmosphérique des polluants acides constitue une menace pour les écosystèmes aquatiques sensibles (ALTSHULLER ET MCBEAN, 1979, 1980; HARVEY ET AL., 1981; États-Unis – Canada 1983). L'évaluation de l'importance de ce phénomène d'acidification pose cependant de grands problèmes logistiques en raison de l'immensité du territoire menacé, du grand nombre de cours d'eau qu'on y retrouve et des particularités régionales qui modifient la réponse de ces milieux. Devant la complexité du problème et les coûts extrêmement élevés d'échantillonnage pour disposer d'une connaissance suffisante de l'étendue du phénomène, on a rapidement senti le besoin de disposer d'un outil de gestion permettant d'évaluer la sensibilité à l'acidification au niveau de l'écosystème aquatique (COWELL ET AL., 1981; ALTSHULLER et MCBEAN, 1980). On a assisté d'abord à l'établissement d'échelles de sensibilité basées principalement sur un seul facteur. Par exemple, on a évalué individuellement l'influence de la géologie (SHILTS, 1981; HENDREY et AL., 1980; NORTON, 1980), des matériaux de surface (SHILTS, 1981), des sols (MCFEE, 1980; WANG et COOTE, 1980) et de la végétation (KLOPATEK et al., 1980; ROBITAILLE, 1980). Par la suite, les efforts de recherche ont porté plutôt sur l'intégration de l'ensemble des facteurs du milieu terrestre à l'établissement d'une échelle valable de vulnérabilité écologique (États-Unis – Canada, 1983). Dans cette optique, des études plus spécifiques ont été amorcées récemment dans le but d'identifier des relations entre la qualité physico-chimique du milieu aquatique et l'ensemble des caractéristiques biophysiques des bassins versants (ROBERTS et al., 1982; RAPP et al., 1985; LOUCKS et al., 1986; HUNSAKER et al. 1986; LACHANCE et GAMACHE, 1985). L'approche généralement retenue a consisté à examiner les corrélations paramétriques ou non paramétriques entre chaque paire de variables et à identifier celles qui sont les plus significatives et pour lesquelles une relation de cause à effet est probable. Pour certaines études, notamment celle de HUNSAKER et al. (1986) dans l'état de New York, de RAPP et al. (1985) dans l'état du Minnesota et celle de ROBERTS et al. (1982) en Suède, on a construit un modèle de régression linéaire multiple permettant de prédire certaines variables physico-chimiques à partir de caractéristiques du bassin versant.

La présente étude s'inscrit dans le même axe de recherche que celui des études précédentes. Cette étude consiste d'une part à analyser la variabilité spatiale de la qualité de l'eau d'un ensemble de lacs d'une région du Québec et d'autre part à rechercher parmi un ensemble de caractéristiques du bassin versant les facteurs explicatifs de cette variabilité spatiale.

L'approche retenue pour cette étude diffère cependant de celle des études précédentes par les méthodes statistiques utilisées qui permettent à la fois une description de la variabilité spatiale d'un ensemble de variables physico-chimiques, d'identifier des relations entre des variables de qualité de l'eau et n'importe quelle type de variables

Ecole d'été sur l'analyse des données, juin 1988,
UQAM, SCAD-CIRADE.

(qualitatives, quantitatives ou ordinales) caractéristiques du bassin versant et de construire un modèle de prévision de certaines variables de qualité de l'eau à partir de quelques variables caractérisant le milieu environnant.

L'analyse de la variabilité spatiale de la qualité de l'eau est effectuée par l'emploi combiné de l'analyse factorielle des correspondances (BENZÉCRI et coll., 1976; LEBART et al., 1979) et de la classification ascendante hiérarchique (JAMBU, 1976; JAMBU ET LEBEAUX, 1983). La recherche des facteurs explicatifs de la variabilité spatiale est réalisée en utilisant l'analyse des correspondances multiples (BENZÉCRI, 1977; CAZES, 1980) sur des tableaux croisant des modalités des variables physico-chimiques avec celles des variables décrivant le milieu environnant. Pour la prédiction des variables de qualité de l'eau, on construit un modèle de régression linéaire multiple à partir des facteurs issus de l'analyse des correspondances du tableau croisé (CAZES, 1976; 1980; 1983).

2 Données

La région du Québec qui a été inventoriée est située dans la partie sud-est du bouclier canadien entre la rivière des Outaouais et la rivière Saguenay. Cette région correspond à une bande de territoire de 150 km de largeur le long de l'axe du fleuve Saint-Laurent (Figure 4). Les données décrivant la qualité physico-chimique des lacs ont été obtenues lors d'un échantillonnage de 158 lacs effectué par Environnement Canada (Direction des eaux intérieures) au cours de l'hiver 1982. La méthodologie utilisée par Environnement Canada pour la sélection et l'échantillonnage des lacs est décrite dans BOBÉE et al. (1983). Notons que tous ces lacs sont situés à la tête des bassins versants dans des régions relativement peu habitées par l'homme.

Parmi la gamme des variables physico-chimiques mesurées, on retrouve entre autres les cations majeurs (calcium, magnésium, sodium et potassium), les anions majeurs (bicarbonates, sulfates, chlorures, nitrates), le pH, la conductivité, la couleur et quelques métaux traces (aluminium, manganèse, fer). On peut trouver dans BOBÉE et al. (1983) un tableau donnant les méthodes utilisées pour les dosages chimiques ainsi que les limites de détection analytique.

Un grand nombre de caractéristiques biophysiques des bassins versants des 158 lacs ont également fait l'objet d'un inventaire par Environnement Canada (Direction des Terres). Une description complète de l'ensemble de ces données inventoriées est donnée par LACHANCE et GAMACHE (1985). Une première interprétation effectuée par ces auteurs a permis d'éliminer un certain nombre de variables non pertinentes et de regrouper les variables pertinentes en cinq catégories: physiographie, géologie, dépôts meubles, végétation et météorologie.

Dans la physiographie, on inclut la variable de localisation du lac (longitude et latitude), l'altitude du lac, la superficie du bassin versant (SBV), la superficie du lac (SLC), la superficie des tourbières (STB) et la pente du bassin. On a considéré également les rapports de superficies SBV/SLC et SBV/STB, et la position des lacs selon un axe SW-NE (éloignement des sources majeures de pollution).

En ce qui concerne la pente du bassin, on a considéré cinq classes: terrain plat (0 à 2%); pente douce (2 à 10%); pente moyenne (10 à 30%); pente forte (30 à 80%); pente très forte à abrupte (> 80%). Chaque bassin versant a été divisé en parcelles selon le type de pente identifiée par photo-interprétation. Pour chacun des 158 bassins versants, on a calculé la superficie relative occupée par chacune de ces classes de pente.

Les roches présentes sur les bassins versants de la région étudiée ont été identifiées à partir de rapports géologiques et de cartes géologiques à différentes échelles; leur superficie relative occupée sur le bassin versant a également été déterminée. Les différents types de roches identi-

fiés ont par la suite été réunis en dix groupes selon leur altérabilité, leur sensibilité à la charge acide et leur pouvoir tampon. Ces groupes comprennent principalement les roches suivantes:

1) marbre et roche calcosilicatée;

2) anorthosite, gabbro, anorthosite gabbroïque et métagabbro;

3) diorite, diorite à pyroxène et diorite gneissique;

4) migmatites, migmatites à biotite, migmatites à trame de gneiss, paragneiss et granites associées, migmatites à plagioclase et migmatites à hornblende;

5) paragneiss, paragneiss de Grenville, paragneiss et roches ignées associées, paragneiss mixtes à amphibole et paragneiss mixtes;

6) gneiss, gneiss intermédiaires, complexe gneissique, gneiss graphitiques et gneiss avec injection;

7) gneiss à biotite, gneiss à pyroxène, gneiss à amphibole, gneiss à quartz, gneiss à hornblende, gneiss à sillimanite et grenat;

8) mangerite, monzonite, migmatites charnockitiques et monzonite à pyroxène;

9) granite avec pegmatite, gneiss granitiques, granodiorite et granite à biotite;

10) quartzite et paragneiss quartzofeldspathique.

En ce qui concerne les dépôts meubles on a mesuré sur le bassin versant la superficie occupée par différents types de dépôts classés selon leur abondance. Les dépôts qui y sont distingués comprennent:

- affleurement rocheux;
- affleurement rocheux et mince pellicule de till;
- till d'épaisseur inférieure à 0,5 m sur assise rocheuse;
- pellicule de till dont l'épaisseur varie entre 0,5 et 1 m;
- till dont l'épaisseur est plus grande que 1 m;
- dépôts organiques dont l'épaisseur dépasse 30 cm;
- sable et gravier.

La variable végétation comprend la mesure, sur le bassin versant, des superficies occupées par différents types de végétation. Les types de végétation qui ont été répertoriés ont été regroupés en cinq classes: arboré résineux dense, arboré résineux clair, arboré feuillu dense, arboré feuillu clair, présence de coupe forestière ou d'arbustes. On a considéré comme dense un couvert arboré d'une densité supérieure à 60%.

Les variables décrivant la météorologie sont la précipitation moyenne annuelle (1931-1960), la concentration de sulfates dans les précipitations, ainsi que le dépôt annuel humide de sulfates pour les années 1982-1983. Les données relatives aux sulfates dans les précipitations ont été fournies par le réseau d'échantillonnage des précipitations du Québec (GRIMARD, 1984). Pour chacun des bassins versants de la région d'étude, la concentration de sulfates et le dépôt de sulfates des précipitations ont été interpolées à partir des iso-lignes de valeurs moyennes annuelles tracées par PARADIS (1985) sur l'ensemble du Québec méridional pour la période 1982-1983.

3 Analyse de la variabilité spatiale de la qualité de l'eau

L'analyse qui est effectuée dans cette section a pour but de décrire globalement les variations spatiales des paramètres physico-chimiques reconnus comme indicateurs du processus d'acidification et de définir sur la région d'étude des zones homogènes correspondant à des régions plus ou moins sensibles à l'acidification par les précipitations acides.

3.1 Choix des variables

Parmi les variables physico-chimiques mesurées, on en a retenu quatre en raison de l'importance de l'information qu'elles fournissent sur le processus d'acidification.

- le pH, qui caractérise le niveau d'acidité du lac. On considère généralement qu'un lac est acide lorsque le pH est inférieur à 5,5.

- l'alcalinité, qui détermine la capacité du bassin versant à neutraliser les apports acides atmosphériques. Une valeur d'alcalinité inférieure à 200 µéq/L est considérée comme indicatrice d'une sensibilité élevée à l'acidification (Bobée et al., 1982). Une valeur inférieure à 100 µéq/L reflète une sensibilité extrême à l'acidification.
- les sulfates, qui traduisent le degré d'agression des eaux de surface par les apports atmosphériques des composés de soufre. Dans les bassins versants où la roche en place est constituée généralement de roches granitiques, comme c'est le cas ici, les sulfates que l'on retrouve dans les lacs sont en concentrations directement proportionnelles aux sulfates dans les précipitations.
- la somme des concentrations de calcium et de magnésium, qui représente le niveau de minéralisation des eaux et conséquemment qui indique la capacité du bassin versant à absorber par échange ionique les ions H^+ en provenance des précipitations.

Les autres variables physico-chimiques mesurées lors de l'inventaire des lacs n'ont pas été retenues en raison de leur importance secondaire comme indicateurs du processus d'acidification. En effet, les ions majeurs non retenus (sodium, potassium, chlorures, nitrates) servent principalement à une validation des données par bilan ionique. La couleur est plutôt indicatrice de la présence de matière organique et d'une contribution possible de cette matière organique à une acidification naturelle. Les métaux lourds Al, Fe et Mn ont été mesurés sous leur forme totale, alors que ce sont les formes dissoutes qui ont une importance dans le processus d'acidification. En outre, la forme dissoute de ces métaux varie grandement avec le pH du milieu aquatique.

3.2 Codage disjonctif des variables

Pour l'application de l'analyse factorielle des correspondances (AFC), il est nécessaire que les données physico-chimiques soient transformées en fréquences. Un codage disjonctif complet a donc été effectué. Le codage disjonctif complet est obtenu de la façon suivante. Soit un tableau X de dimensions n x p comprenant en colonnes les p variables et en lignes les n lignes. Le codage disjonctif complet d'une variable j en k classes consiste à remplacer dans le tableau X la colonne j par k colonnes comprenant seulement des 1 ou des 0 selon que la valeur X appartient ou n'appartient pas à la classe correspondante.

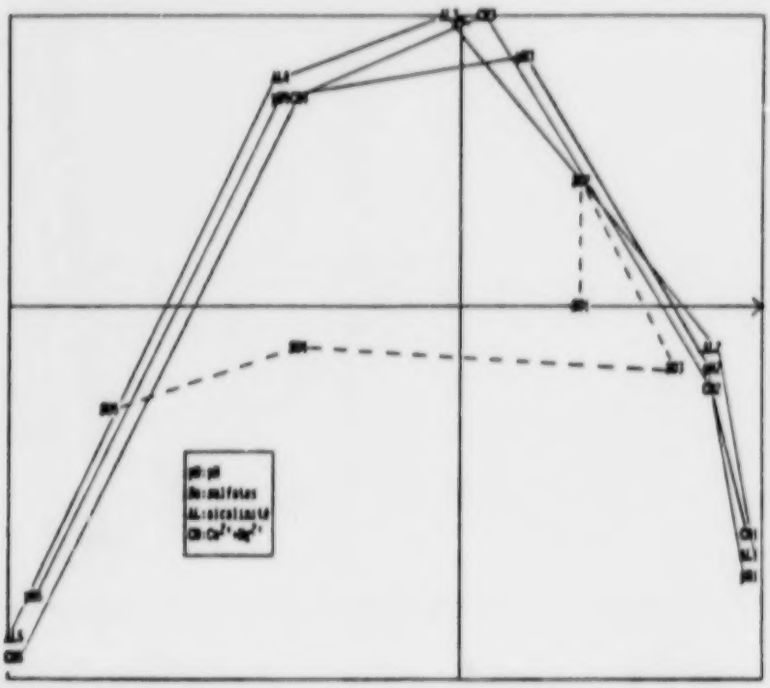
Pour l'analyse qui sera effectuée par l'AFC, on a choisi de considérer cinq classes équiprobables par variable. Le tableau qui sera analysé par l'AFC a donc comme dimensions 158 lignes correspondant aux 158 lacs et 20 colonnes correspondant aux quatre variables séparées en cinq classes.

3.3 Analyse des points variables

Cette application de l'AFC permet de représenter dans un espace de dimensions réduites les variables considérées. Les trois premiers axes principaux expliquent respectivement 20,1%, 15,1%, et 9,7% de la variabilité totale pour un total de 44,9%. L'examen de la position des points variables dans le plan des axes 1-2 et 1-3 (Figure 1) montre que les variables pH, alcalinité et Ca + Mg ont une variation identique alors que la variable sulfates a un comportement différent. On obtient en effet, une similitude de parcours pour ces trois variables en reliant par des droites les points représentant des classes voisines. En interprétant la position des points représentant les classes des sulfates (spécialement les points SU5 et SU4), on en déduit que les valeurs élevées de sulfates correspondent généralement à des valeurs élevées des trois autres variables; par contre, on retrouverait des valeurs intermédiaires de sulfates là où les trois autres variables prennent des valeurs faibles.

Figure 1 a): Représentation dans le plan des axes 1-2 des points variables dans l'analyse de la variabilité spatiale de la qualité de l'eau des lacs

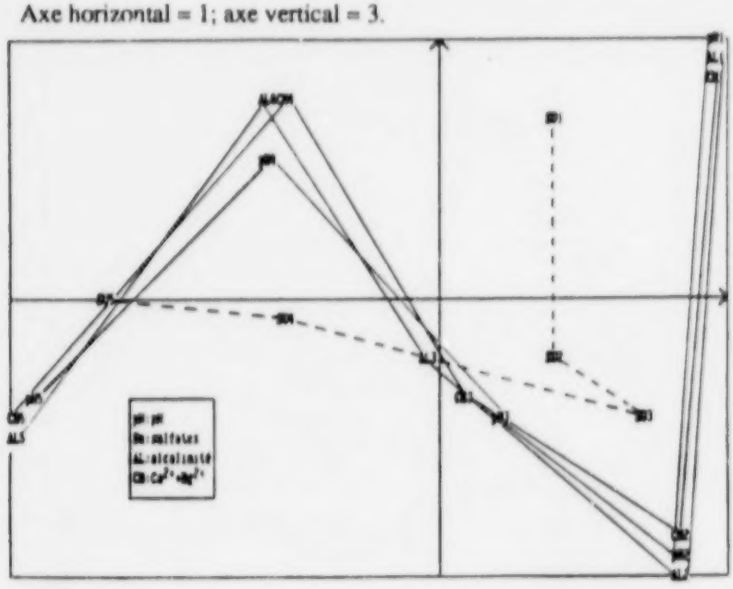
Axe horizontal = 1; axe vertical = 2.



Les points représentant les classes des variables alcalinité, pH et Ca + Mg sont répartis selon une forme parabolique. Rappelons que ce type de parcours (effet Guttman) est obtenu quand le phénomène est principalement unidimensionnel (BENZÉCRI et coll., 1976). Dans ce cas, le deuxième facteur est une fonction du deuxième degré du premier facteur, le troisième facteur est une fonction cubique du premier, etc... On retrouve donc sur l'axe 1 une opposition entre les classes extrêmes alors que sur les axes successifs (2 et 3 en particulier), ce sont surtout les classes intermédiaires qui se retouvent sur les extrémités des axes. Ce type de représentation indique donc que le phénomène spatial, décrit par les variables alcalinité, pH et Ca + Mg, présente une certaine continuité; sur la région d'étude, la variation spatiale de la minéralisation des eaux expliquerait ce phénomène de continuité spatiale.

En examinant la qualité de la représentation des points variables sur les quatre premiers axes principaux (LACHANCE, 1988), on a constaté que les points variables pour les classes 1 à 5 du pH, de l'alcalinité et du calcium + magnésium sont généralement bien représentés sur ces axes (67% en moyenne). Les sulfates sont par contre bien représentés en considérant huit facteurs; la variabilité des sulfates n'est expliquée vraiment que sur les axes 5 à 8, comme l'a montré le calcul des contributions relatives cumulées sur quatre et sur huit axes (LACHANCE, 1988).

Figure 1 b): Représentation dans le plan des axes 1-3 des points variables dans l'analyse de la variabilité spatiale de la qualité de l'eau des lacs



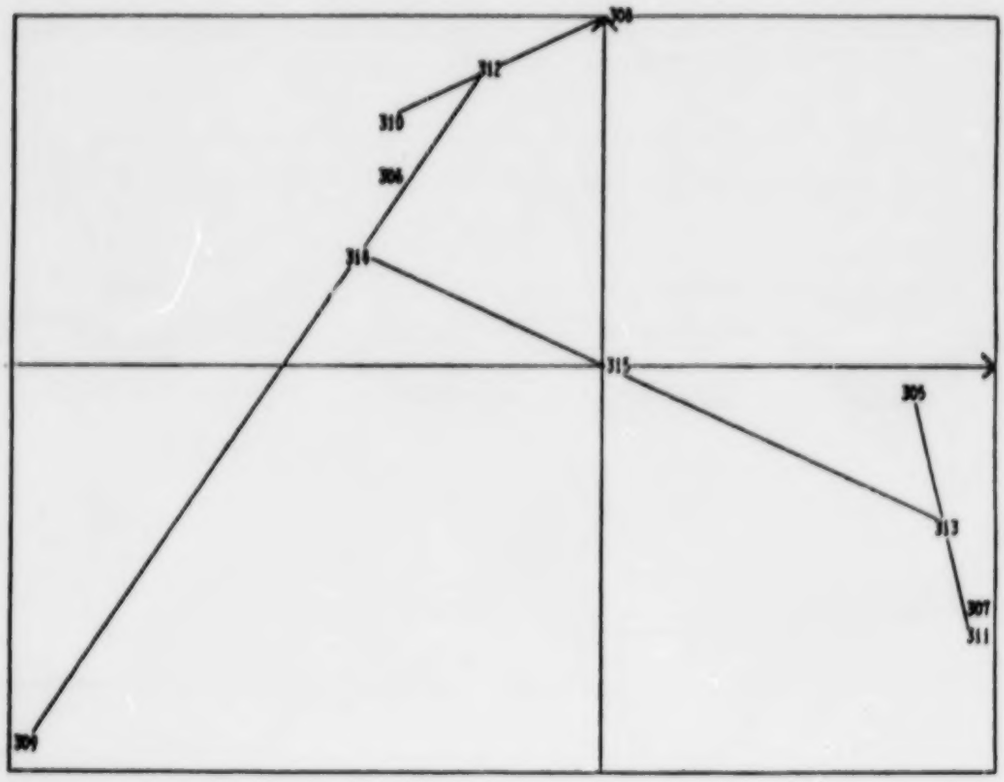
3.4 Analyse des points observations

L'analyse et l'interprétation des points observations dans l'espace factoriel sont grandement facilitées en utilisant une méthode de groupement telle que la classification ascendante hiérarchique (CAH). Cette méthode peut être appliquée soit sur un tableau de correspondances, soit sur un tableau de coordonnées factorielles. Dans ce cas-ci, on a choisi d'appliquer la CAH sur le tableau des facteurs, car les résultats de la CAH ainsi obtenus permettent d'enrichir considérablement l'interprétation des résultats de l'AFC. D'autre part, le fait de retenir un nombre réduit de facteurs permet d'éliminer du tableau de données un bruit de fond attribuable aux fluctuations d'échantillonnage ou à l'imprécision des mesures. On a observé précédemment qu'avec huit facteurs l'ensemble des points variables étaient bien représentés. On a donc appliqué la CAH sur le tableau des huit premiers facteurs obtenus de l'AFC; la représentation de la hiérarchie supérieure du classement est fournie à la figure 2a.

Pour identifier les noeuds de la hiérarchie, on a l'habitude de les numéroter de n = N+1 à n = 2N-1 selon un indice d'inertie croissant (n) où N est le nombre d'observations à classer. Dans ce cas-ci, les noeuds de l'arbre hiérarchique sont numérotés de 159 à 315. Le taux d'inertie associée au noeud 315, formé par la réunion des classes 313 et 314, correspond au maximum de dispersion inter-groupe. La classe correspondant au noeud 315 représente le nuage de l'ensemble des 158 points. En coupant l'arbre hiérarchique à un certain niveau, on a obtenu une partition des lacs en cinq classes, soit les classes de lacs associées aux noeuds 311, 305, 309, 308 et 310 (Figure 2a).

L'interprétation de la signification du classement obtenu est facilitée en représentant les centres de gravité des classes de la hiérarchie dans l'espace factoriel (Figure 2b). On voit que la première division des lacs (séparation de la classe 315 en 314 et 313) se fait selon l'axe 1, alors que la seconde division, obtenue en considérant la séparation de la classe 314 en ses deux constituants 309 et 312, est plutôt dans la direction de l'axe 2. Pour interpréter la division de la classe 313 en 305 et 311 et celle de la classe 312 en 308 et 310, il faut considérer les axes 3 et 4. Il est possible d'évaluer plus exactement la signification du classement dans l'espace factoriel en considérant la contribution relative de chaque facteur à chaque noeud. Le détail de ces calculs peut être trouvé dans LACHANCE (1988).

Figure 2 a): Classification hiérarchique des lacs de la région d'étude: arbre hiérarchique.



Une caractérisation des groupes peut être obtenue en examinant dans l'espace factoriel les proximités entre points variables et groupes de points stations (Figure 3a). En effet, grâce à une des propriétés de l'AFC (effet barycentrique), une proximité plus ou moins grande entre des points variables et un groupe de points stations implique une contribution plus ou moins grande de ces points variables à la formation du groupe de points stations. La figure 3a fournit la représentation des points variables et des centres de gravité des cinq groupes de points lacs dans le plan des axes 1-2. Pour une simplification de la

notation, les groupes de lacs ont été désignés par les lettres A à E de la façon suivante:

311	305	308	310	309
A	B	C	D	E

Par l'examen des proximités entre points variables et groupes de lacs, on peut ainsi caractériser chacun des cinq groupes. On peut ainsi distinguer que le groupe E, par exemple, est caractérisé principalement par de fortes valeurs de pH (pH5), d'alcalinité (AL5) et de calcium + magnésium (CM5) et à un degré moindre par de fortes valeurs de sulfates (SU5). Il est possible de caractériser d'une façon plus objective les groupes de lacs en calculant la contribution de chaque point variable à la position du centre de gravité du groupe dans l'espace factoriel. Les détails de ces calculs sont fournis par LACHANCE (1988).

Figure 2 b): Classification hiérarchique des lacs de la région d'étude: représentation des centre de gravité des classe dans le plan factoriel 1-2.

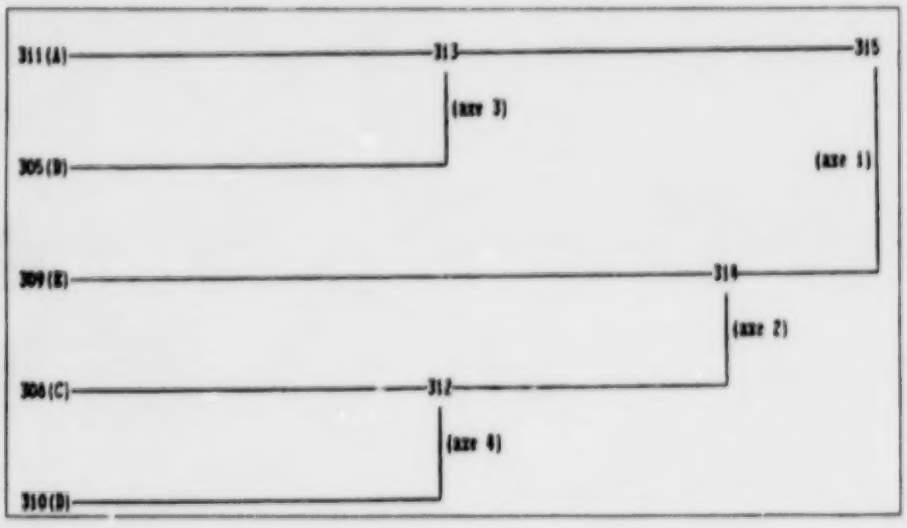
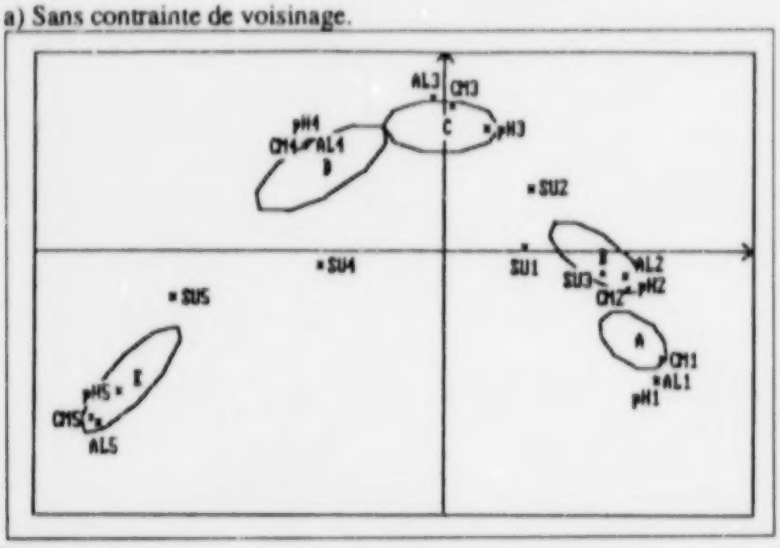


Figure 3 Représentation simultanée des points variables et des nuages de points lacs obtenus en appliquant la CAH avec différentes contraintes de voisinage.

a) Sans contrainte de voisinage.



b) Avec une faible contrainte de voisinage.

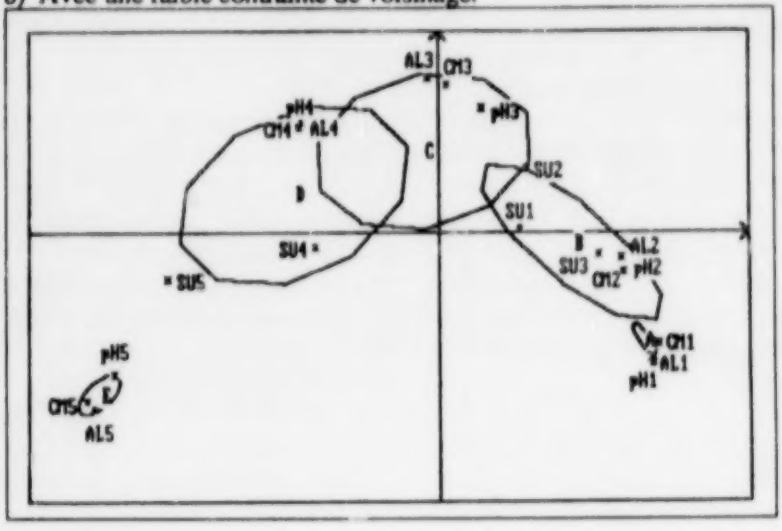
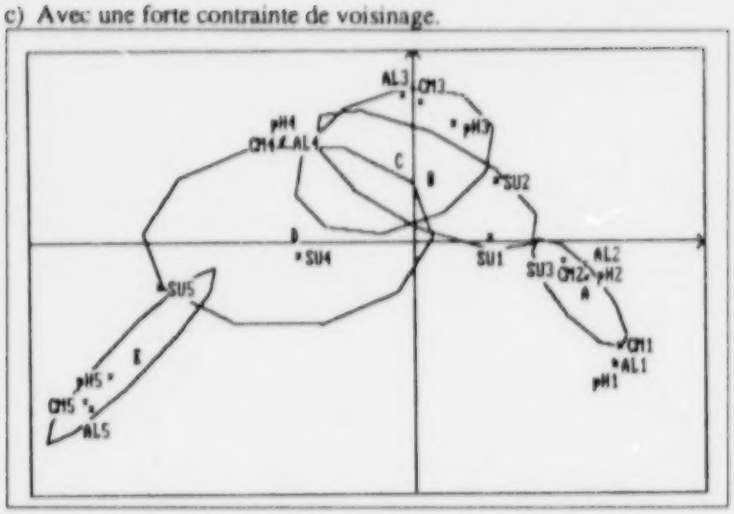


Figure 3 (suite).

c) Avec une forte contrainte de voisinage.



De l'examen de cette représentation, on voit clairement que ce sont les classes des variables alcalinité, calcium + magnésium et pH qui caractérisent chacun des cinq groupes. À l'exception du point SU5 pour le groupe E, la variable sulfates joue un rôle moins déterminant. Ce résultat est concordant avec le fait que la partition des lacs en cinq groupes est bien expliquée par les quatre premiers facteurs sur lesquels les variables pH, Ca + Mg et alcalinité sont bien représentées, alors que les sulfates sont plutôt représentés sur les axes supérieurs.

Pour avoir une idée de la dispersion du nuage correspondant à chacun des cinq groupes de lacs, on a procédé au calcul des ellipsoïdes d'inertie selon la méthode décrite par GRELET et LEBEAUX (1980). L'examen de la grandeur des ellipses d'inertie représentées dans le plan des axes 1-2 (Figure 3a) permet de constater que la dispersion de chacun des nuages de points est comparable. L'absence de chevauchement des ellipses confirme l'observation que ces cinq classes de lacs forment des groupes bien distincts l'un de l'autre.

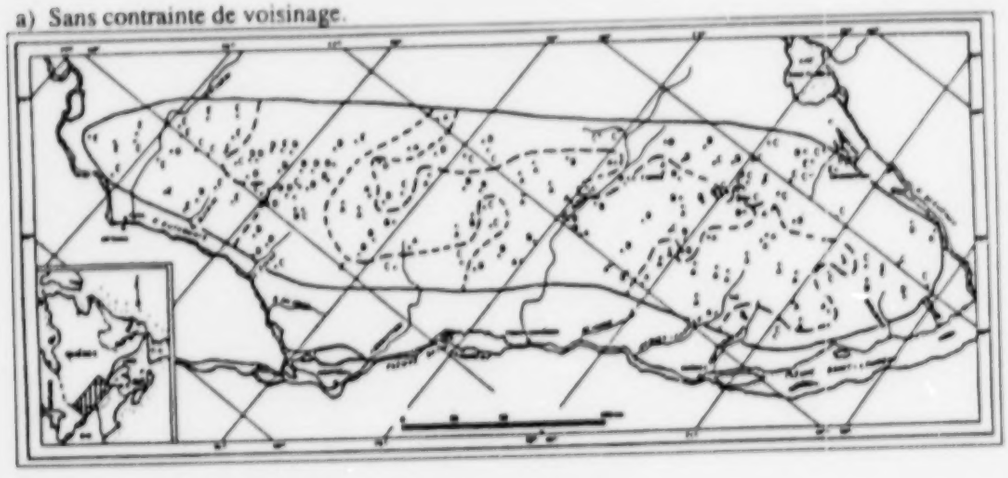
L'examen de la répartition géographique sur la région d'étude des cinq groupes de lacs (Figure 4a) permet de compléter la description de la structure spatiale de la région. La partie sud-ouest présente une forte densité de lacs appartenant au groupe E, tandis que dans la partie nord-est (plus précisément, dans la réserve des Laurentides au nord de Québec), on distingue une région appartenant principalement au groupe A. Les zones de transition, malgré une plus grande diversité, comprennent cependant des groupes dont les caractéristiques sont voisines.

Dans une étude antérieure, Bobée et al. (1983) ont délimité sur cette région cinq zones géographiques (Figure 4a) en se basant sur les résultats obtenus de cette classification et en utilisant en outre de l'information sur les caractéristiques physiographiques du territoire. Cette délimitation quelque peu subjective a été effectuée dans le cadre de l'élaboration et de la mise en

place au Québec d'un réseau de surveillance de l'acidification des eaux lacustres (Bobée et al., 1983).

Figure 4: Regroupement des lacs en cinq zones en appliquant la CAH

a) Sans contrainte de voisinage.



b) Avec une faible contrainte de voisinage.

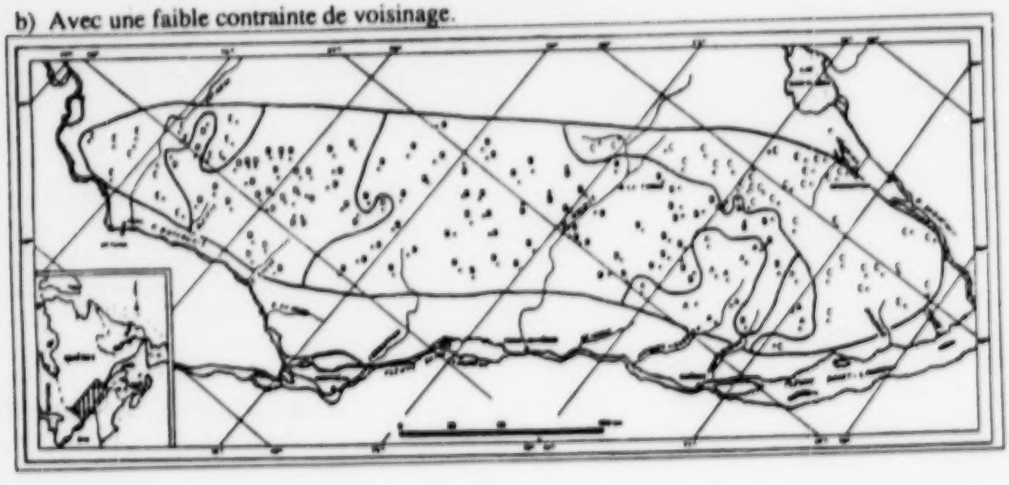
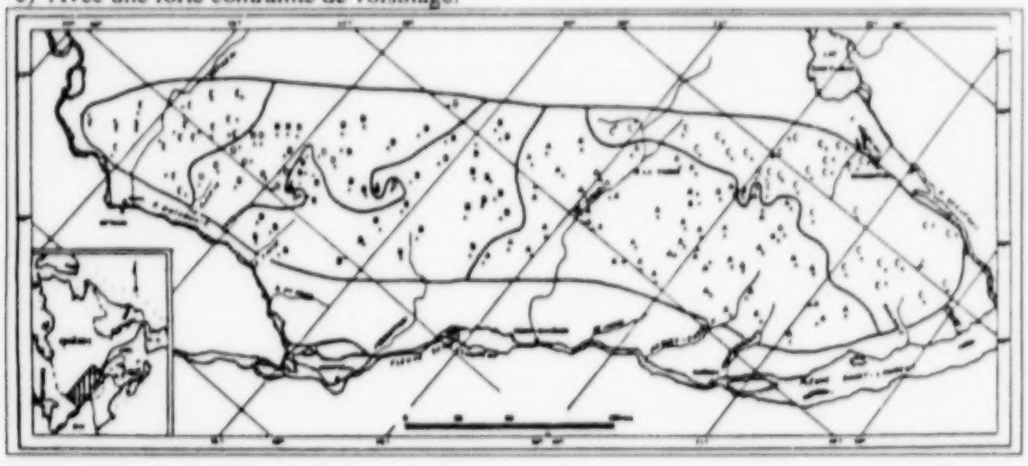


Figure 4c): (suite).

c) Avec une forte contrainte de voisinage.



3.5 Analyse avec une contrainte de voisinage

Afin d'obtenir une zonation plus objective, une contrainte de contiguïté a été imposée à la méthode de classification. Cette technique, mise au point par LEBART (1978), permet d'obtenir automatiquement une partition des stations en classes connexes géographiquement, les regroupements étant effectués à partir des similitudes de caractéristiques des variables considérées. L'algorithme de la méthode utilisée est essentiellement le même que pour la CAH. On procède d'abord au calcul des distances entre chaque paire possible de points. À partir des coordonnées des points sur les premiers axes factoriels, la distance euclidienne est celle qui est habituellement calculée. On recherche ensuite parmi les n objets à classer les deux objets les plus proches (donc à distance minimale). Cependant, au lieu de chercher les deux objets les plus proches parmi tous les couples possibles, on se limite à les chercher parmi seulement tous les couples géographiquement voisins. Après avoir agrégé les deux objets en un nouvel objet, on procède au calcul des distances entre ce nouvel objet et tous les autres. La seconde étape est alors analogue à la première, mais avec (n-1) objets. Le processus continue jusqu'à ce qu'il ne reste plus qu'un seul objet.

L'utilisation de la méthode nécessite donc de préparer une matrice de contiguïté donnant pour chaque lac les voisins les plus proches. Des polygones de Thiessen ont alors été tracés pour chaque lac (LACHANCE, 1988). Deux notions de contiguïté ont ensuite été définies:

- Un lac est reconnu contigu à un autre lac lorsque ces deux lacs ont en commun un côté de leur polygone. Le nombre de lacs contigus à un lac correspond au nombre de côtés du polygone tracé pour ce lac.

- Un lac est reconnu contigu à un autre lac lorsque ces deux lacs ont en commun soit un coté, soit un sommet de leur polygone. Cette notion de contiguïté a donc un sens plus large.

Lorsqu'on distingue, dans le texte qui suit, une forte ou une faible contrainte de contiguïté, on réfère donc à la première ou à la seconde situation décrite ci-dessus.

Pour l'interprétation des nouvelles classifications obtenues, on compare les groupes obtenus en représentant dans l'espace factoriel leur centre de gravité et leur ellipsoïde d'inertie (Figures 3b et 3c) et on examine sur la région d'étude les zones délimitées par ces groupes (Figures 4b et 4c). Comme précédemment, chacune des cinq classes a été désignée par les lettres A à E. Dans l'attribution des lettres aux classes, on a tenté de leur conserver, autant que possible, la même signification que pour l'analyse effectuée sans contrainte de voisinage.

Avec une forte contrainte de voisinage, on a obtenu une augmentation importante de la dispersion intra-classe spécialement pour les classes B, C et D (Figure 3c). Le chevauchement des nuages de points rend difficile l'identification des caractères propres à chacune des classes. Avec une contrainte moins forte, le chevauchement des groupes existe encore (B, C et D), mais est plus faible (Figure 3b). La plus grande différence, cependant, se situe au niveau des groupes A et E; la diminution de la contrainte de contiguïté a entraîné une plus grande homogénéité des classes A et E.

Les cinq zones géographiques obtenues avec les contraintes de voisinage correspondent grossièrement aux anciennes zones, mais sont maintenant bien délimitées (Figures 4b et 4c). Avec une forte contrainte de contiguïté, les cinq zones délimitées sont bien équilibrées au niveau de la répartition des effectifs (Figure 4c). La forte dispersion intra-groupe est expliquée par le regroupement forcé dans une zone de lacs possédant des caractéristiques fort différentes de celles des lacs voisins. La diminution de la contrainte de voisinage se fait principalement sentir au niveau d'une diminution de la superficie des zones A et E et, d'une augmentation importante de celle de la zone B (Figure 4b). La nouvelle zone A ne comprend maintenant plus que 12 lacs et est confinée à la partie sud de la réserve des Laurentides (nord de Québec).

On peut retenir de cette dernière analyse que le fait de rendre moins sévère l'application de la contrainte de contiguïté permet d'améliorer la définition des zones de qualité. Les zones géographiques obtenues par cette technique peuvent être mieux caractérisées en termes de niveaux ou classes de valeurs des variables considérées.

4 Recherche des facteurs explicatifs de la variabilité spatiale

Dans cette section, on montre comment l'application de techniques d'analyse des correspondances multiples (ACM) permet de procéder à l'examen des liens pouvant exister entre les variables physico-chimiques considérées précédemment et un ensemble de variables biophysiques (décrivant la physiographie, la géologie, la végétation, les sols, etc...) mesurées sur le bassin versant des lacs de la région. Cette analyse des relations entre la qualité de l'eau et les caractéristiques du territoire conduira à l'élaboration d'un modèle de régression permettant de prédire une variable de qualité de l'eau à partir de quelques facteurs du bassin versant.

Pour l'analyse des relations entre les deux groupes de variables, soit les variables de qualité de l'eau et les variables du bassin versant, on applique l'analyse des correspondances sur un tableau croisant les modalités d'un groupe de variables K ($K \in L$) avec celles d'un autre groupe de variables K' ($K' \in L'$). Les propriétés des facteurs issus de l'analyse de ce tableau sont analogues à celles obtenues de l'analyse du tableau de Burt (CAZES, 1980).

L'identification des liens entre variables du groupe L et celles du groupe L' est rendue possible en représentant simultanément sur les plans factoriels les points j de l'ensemble L et les points j' de l'ensemble L'. En effet, en vertu de la propriété barycentrique de l'AFC, chaque

point j de L (resp. j' de L') est à une constante près le centre de gravité des points j' de L' (resp. j de L).

4.1 Données considérées

Les variables que l'on cherche à expliquer sont les quatre variables physico-chimiques qui ont servi à la détermination de cinq zones homogènes de qualité. Ces variables sont l'alcalinité, les sulfates, le pH et le calcium + magnésium. Les variables du bassin versant que l'on tente de relier avec les précédentes appartiennent à chacune des cinq catégories suivantes: physiographie, géologie, dépôts meubles, végétation et météorologie. Ces variables ont été décrites précédemment.

L'ensemble I des observations sur lesquelles on cherche à identifier des relations entre variables physico-chimiques et variables du milieu peut être constitué de l'ensemble de la région d'étude ou de un ou plusieurs sous-groupes déterminés par l'application de la CAH avec ou sans contrainte de voisinage.

Tenant compte d'un ensemble de considérations, l'ensemble I qui a été retenu pour l'analyse qui est présentée ici est constitué de l'ensemble des groupes A, B, C et D (Figure 3a) issus de la CAH appliquée sans contrainte de voisinage. En fait, on exclut les 24 lacs du groupe E.

Les lacs du groupe E sont principalement situés dans la partie sud-ouest de la région d'étude. Cette zone qui est caractérisée par les plus fortes valeurs de pH, d'alcalinité, de Ca + Mg et de sulfates, présente des caractéristiques physiographiques et géologiques fort différentes de celles retrouvées sur l'ensemble de la région. En particulier, on y retrouve des roches calcaires (calcaire cristallin) dont la présence contribue à des teneurs très élevées de Ca + Mg et d'alcalinité dans les eaux. En excluant le groupe E de l'ensemble I, on s'assure d'une plus grande homogénéité dans les données, sans toutefois empêcher de dégager des relations globales au niveau de l'ensemble de la région d'étude.

L'obtention du tableau, sur lequel on applique l'ACM nécessite de franchir deux étapes de préparation des données: le codage binaire des données et leur transformation en tableaux de contingence.

4.2 Codage binaire des données

Le tableau initial décrivant les mesures des quatre variables physico-chimiques sur les lacs et les douze variables biophysiques sur le bassin versant des lacs est hétérogène en ce sens qu'il comporte des variables quantitatives et des variables qualitatives. Un codage disjonctif complet est donc effectué sur les variables quantitatives. Le choix du nombre J_q de classes par variables est toujours un problème délicat en analyse de données. De façon générale, le codage disjonctif résulte toujours en une certaine perte d'information et cette perte d'information est d'autant plus grande que le nombre de classes est faible. Un codage flou (réparti sur deux classes voisines) est parfois utilisé pour minimiser cette perte d'information (GALLEGO, 1982).

En ce qui concerne les variables du milieu (considérées comme variables explicatives), un choix de cinq classes d'effectifs semblables a été retenu compte tenu du nombre d'observations (134) et du compromis à faire entre la perte d'information résultant du codage et le sens physique des classes retenues. En ce qui concerne les variables qualitatives telles que la pente du bassin, la géologie, la végétation et le type de sols, le nombre de classes choisies correspond au nombre de modalités retenues (4 pour la pente, 5 pour les autres). Pour ces variables, les effectifs par classe peuvent parfois s'éloigner de l'équi-probabilité. Pour l'ensemble des 12 variables explicatives, le nombre total de classes s'élève à 59.

Pour les variables physico-chimiques (considérées comme variables à expliquer), on a tenté

d'effectuer un découpage plus fin en dix classes d'effectifs semblables. Dans un souci de cohérence, on a conservé les seuils établis précédemment pour le découpage de ces variables en cinq classes. Pour effectuer ce découpage plus fin, on ajoute donc des seuils intermédiaires permettant de subdiviser en deux les intervalles de variation définis précédemment. À cause de l'exclusion des 24 lacs du groupe E, on constate qu'on ne peut obtenir dix classes par variable, puisque ces lacs constituent principalement l'effectif des classes 9 et 10. On a donc conservé 8 classes pour l'alcalinité et le calcium + magnésium et 9 classes pour les sulfates et le pH.

4.3 Calcul des tableaux de contingence

Le tableau de contingence C (de dimensions $J_1 \times J_2$) est obtenu en comptant le nombre d'occurrences simultanées des modalités de l'ensemble J_1 et de celles de l'ensemble J_2. Puisque l'analyse des correspondances a été appliquée sur chacun des tableaux croisant une variable physico-chimique et un ensemble de variables du milieu, le tableau analysé a comme dimensions 59 colonnes et 8 ou 9 lignes (dépendant de la variable physico-chimique considérée).

Relations alcalinité - variables du milieu

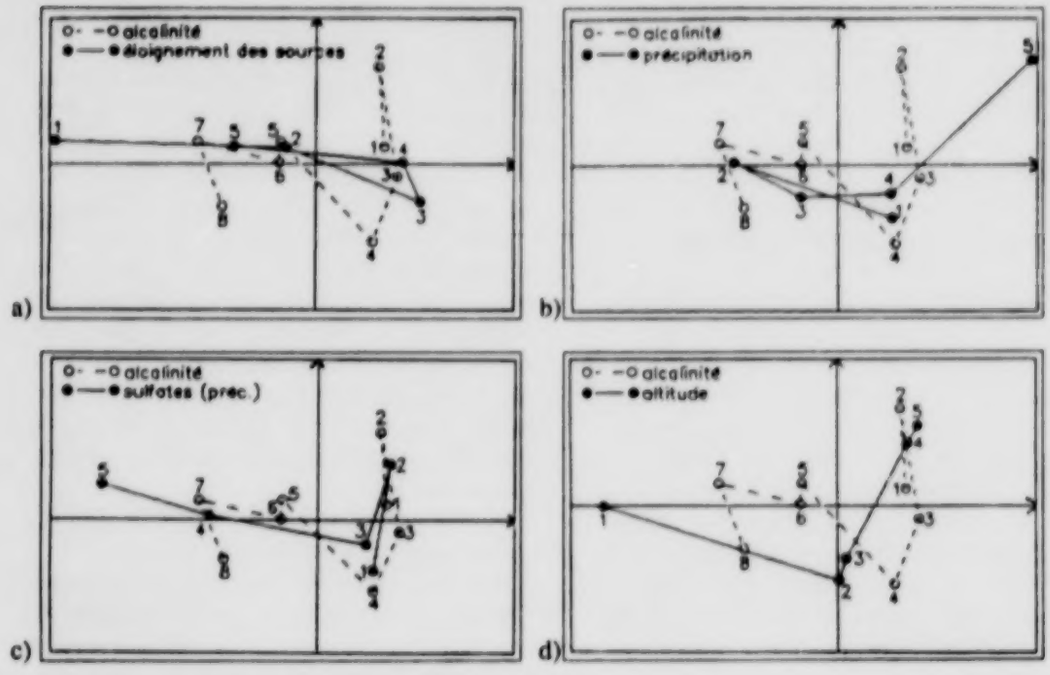
L'analyse des correspondances du tableau croisant les classes de l'alcalinité avec celles des variables du milieu permet d'obtenir une représentation dans un espace factoriel de l'ensemble des points i (8 classes de l'alcalinité) et de l'ensemble des points j (59 classes des variables du milieu). Les quatre premiers axes factoriels expliquent 39,2%, 18,2%, 13,4% et 11,3% de l'inertie totale du nuage de points. On voit donc qu'avec deux axes, l'essentiel de la variabilité (57,4%) contenue dans le tableau de contingence est expliqué.

L'analyse des relations entre l'alcalinité et les variables du milieu est effectuée en projetant d'abord dans le plan des axes 1 et 2 les points de l'ensemble I (8 classes de l'alcalinité) et les points de l'ensemble J (59 classes au total). On relie ensuite par une droite les points représentant les classes voisines de chaque variable (lorsque le sens physique le permet). Enfin, on identifie les liaisons entre l'alcalinité et chacune des variables explicatives en examinant les similitudes de parcours et en interprétant les proximités entre points. On peut trouver dans LACHANCE (1988) une interprétation détaillée des relations entre l'alcalinité et chacune des variables du milieu. Dans le texte qui suit, on décrit les résultats obtenus en ce qui concerne les variables "éloignement des sources", "hauteur de précipitation", "concentration de sulfates dans les précipitations" et "altitude du lac" (Figures 5a à 5d). Pour compléter l'interprétation de ces graphiques, on présente des coefficients de corrélation de Pearson et Spearman calculés entre l'alcalinité et quelques-unes de ces variables (Tableau 1).

Positionnement de l'alcalinité

Sur l'axe 1, les points représentant les classes de l'alcalinité se projettent selon un ordre généralement décroissant avec une petite inversion pour la classe 7 et les classes 1 et 2. L'axe 1 exprime donc une décroissance générale des valeurs d'alcalinité (et conséquemment de la sensibilité à l'acidification) en se déplaçant de gauche à droite sur l'axe. Le phénomène responsable de la variation générale de l'alcalinité n'apparaît pas s'appliquer lorsque les valeurs d'alcalinité sont faibles (classes 1 et 2 notamment). L'opposition sur l'axe 2 de la classe 4 aux classes 1 et 2 indique que sous un certain seuil, la variation de l'alcalinité est due soit à un autre phénomène complètement indépendant du premier, soit à une variation aléatoire provenant d'une imprécision des mesures lorsque les valeurs sont très faibles.

Figure 5 a)-d): Représentation du plan des axes 1 (horizontal) et 2 (vertical) montrant la relation entre certaines variables du milieu et l'alcalinité dans les lacs.



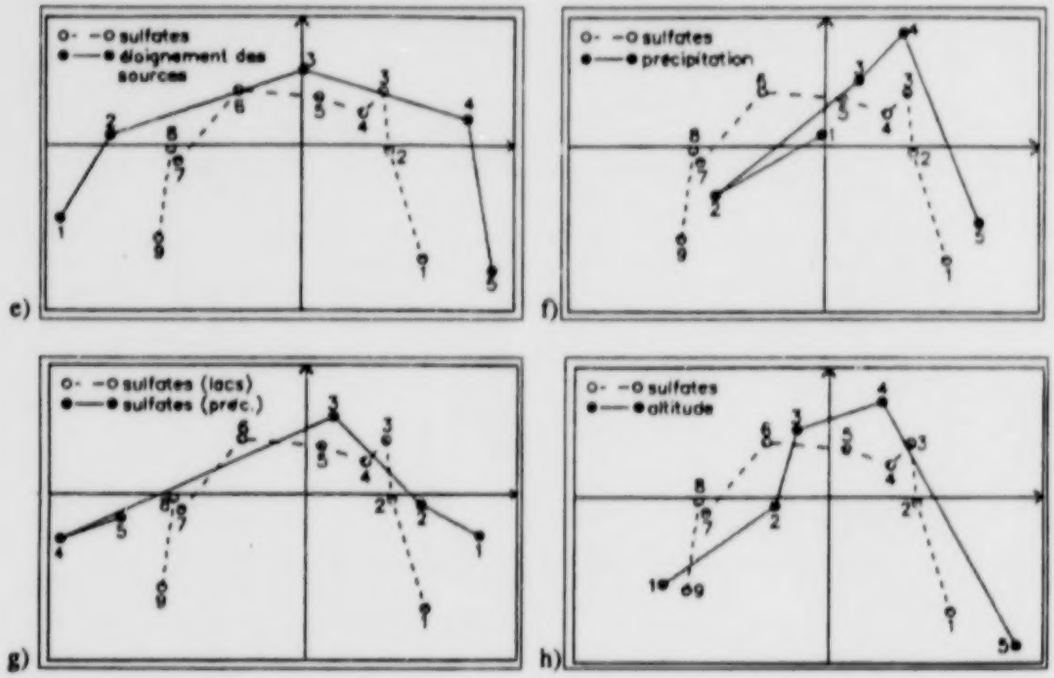
Alcalinité - éloignement des sources

Les points représentant les classes 1 à 3 de la variable "éloignement des sources" se répartissent sur l'axe 1 (Figure 5a) selon un ordre croissant (direction SW-NE) indiquant que sur la majeure partie de la région d'étude (partie sud-ouest et partie centrale), l'alcalinité diminue avec l'éloignement des sources de pollution. Par contre, dans la partie nord-est (classe 5), on retrouve des valeurs d'alcalinité plutôt élevée. Cette inversion est responsable du coefficient de corrélation faiblement significatif (approximativement -0,20; cf. Tableau 1) avec l'alcalinité.

Alcalinité - sulfates dans les précipitations

La variation des sulfates dans les précipitations s'effectue dans le même sens que celle de l'alcalinité dans les lacs (Figure 5b). Les points représentatifs des classes de sulfates se projettent selon un ordre décroissant sur l'axe 1, à l'exception de la classe 1 qui présente une petite inversion. Les valeurs très faibles de sulfates dans les précipitations sont plutôt associées aux valeurs intermédiaires d'alcalinité (classes 3 et 4). Le coefficient de corrélation calculé entre les deux variables (0,30 à 0,33; c. f. Tableau 1) confirme que la relation est plutôt faible, mais présente.

Figure 5 e)-h): Représentation du plan des axes 1 (horizontal) et 2 (vertical) montrant la relation entre certaines variables du milieu et les sulfates dans les lacs.



Alcalinité - hauteur de précipitation

On identifie la présence d'un lien entre la précipitation annuelle et l'alcalinité, celle-ci diminuant avec l'augmentation de la quantité de précipitation. Cette relation, cependant, est obtenue seulement lorsque la valeur annuelle dépasse le seuil de 900 mm (classe 1). On observe en effet que tous les points représentant les classes de précipitation, à l'exception de celui pour la classe 1, se projettent sur l'axe 1 selon un ordre croissant (Figure 5c). Les faibles précipitations (classe 1) sont plutôt associées aux classes intermédiaires 3 et 4 de l'alcalinité. Cette inversion se traduit par un coefficient de corrélation relativement faible (-0,37 pour r de Pearson, -0,32 pour r de Spearman; cf. Tableau 1). En effet, si on exclut du calcul de la corrélation les valeurs de précipitation sous le seuil de 900 mm/année, le coefficient augmente à -0,47 (r de Pearson et Spearman).

Cette relation entre l'alcalinité et la précipitation présente un grand intérêt. On peut en déduire que sur cette région, la sensibilité des lacs à l'acidification, telle qu'estimée par l'alcalinité, augmente avec la hauteur de précipitation lorsque celle-ci dépasse un certain seuil.

Alcalinité - altitude du lac

Le parcours tracé dans le plan 1-2 par les classes de l'altitude (Figure 5d) a une certaine forme parabolique. Cette forme parabolique (opposition des classes extrêmes sur l'axe 1 et des classes intermédiaires sur l'axe 2) peut s'expliquer par le fait que l'axe 1 est vraiment un axe

qui exprime les variations de l'altitude (effet Guttman).

Tableau 1: Coefficients de corrélation de Pearson et de Spearman entre des variables physico-chimiques mesurées sur 134 lacs et des caractéristiques biophysiques mesurées sur leurs bassins versants.

	Variables physico-chimiques			
Variables du milieu	alcalinité	sulfates	pH	Ca + Mg
Éloignement des sources	0,21 -0,19	-0,74 -0,77	-0,29 -0,28	-0,46 -0,41
Sulfates dans les précipitations	0,33 0,30	0,71 0,70	0,41 0,40	0,58 0,54
Dépôt atmosphérique de sulfates	0,06 0,15	-0,23 -0,27	0,10 0,17	-0,02 0,01
Précipitation	-0,37 -0,32	-0,33 -0,32	-0,38 -0,32	-0,46 -0,47
Superficie B. V. (SBV)	-0,02 -0,02	0,04 0,02	0,10 0,01	-0,03 0,03
Superficie lac (SLC)	0,03 0,01	-0,11 -0,07	0,07 0,10	-0,02 -0,02
SBV/SLC	-0,08 -0,02	0,22 0,08	-0,22 0,09	-0,02 0,07
Altitude	-0,35 -0,42	-0,47 -0,49	-0,34 -0,43	-0,49 -0,56

Note:

- corrélation de Pearson
- corrélation de Spearman

$|r| > 0,17$ est significativement différent de zéro au niveau de 5%;
$|r| > 0,22$ est significativement différent de zéro au niveau de 1%.

La comparaison des parcours obtenus pour l'alcalinité et l'altitude indique que cette relation est présente, quoique relativement faible comme on peut le constater par les coefficients de corrélation (r de Pearson = 0,35 et r de Spearman = -0,42). L'altitude a depuis longtemps été reconnue comme un facteur d'explication de la minéralisation des eaux et par conséquent de la sensibilité à l'acidification. Cette relation entre l'altitude et la minéralisation des eaux est due à un ensemble de facteurs plus ou moins importants tels que la quantité de précipitation, la nature des minéraux présents, l'abondance et la texture des dépôts de surface, le temps de contact moyen entre l'eau et les minéraux, etc. La relation altitude - minéralisation des eaux peut être plus ou moins forte selon que l'effet résultant de l'ensemble de ces facteurs est plus ou moins important.

Relations sulfates - variables du milieu

L'étude des relations de la variable sulfates

avec les variables du milieu est effectuée en analysant par l'AFC le tableau croisant les classes de sulfates avec les classes des variables du milieu. L'analyse des correspondances du tableau croisant les classes des sulfates avec celles des variables du milieu fournit une représentation factorielle dans laquelle les premiers axes expliquent 42,9%, 20,8%, 10,9% et 7,3% de la variabilité totale. Le plan des axes 1 et 2 (63,7% de la variabilité) fournit donc une bonne représentation de l'essentiel de la variabilité contenue dans le tableau analysé.

L'analyse des relations entre les sulfates et les variables du milieu est effectuée en suivant exactement la même démarche que celle adoptée précédemment avec l'alcalinité.

Positionnement des sulfates

Les points représentant les classes des sulfates dans le plan 1- 2 (Figures 5e à 5h) sont réparties selon une forme parabolique indiquant encore la présence de l'effet Guttman. La variation des sulfates est donc un phénomène principalement unidimensionnel.

Sulfates - éloignement des sources

La relation entre les sulfates et l'éloignement des sources est très forte comme on peut le constater à partir de l'examen de la position dans le plan 1-2 des points représentant les classes de ces deux variables (Figure 5e). Le parcours obtenu pour la variable "éloignement des sources" a une forme parabolique qui ressemble beaucoup à celle obtenue pour les sulfates. Le coefficient de corrélation entre ces deux variables est d'ailleurs assez élevé (-0,74 pour r de Pearson et -0,77 pour r de Spearman; cf. Tableau 1).

Sulfates dans les lacs - sulfates dans les précipitations

Sur la région d'étude, la concentration de sulfates dans les précipitations est fortement corrélée ($\hat{\beta}$ = 0,70; cf. Tableau 1) avec celle dans les lacs. Sur le plan factoriel, les deux variables présentent une similitude de parcours (Figure 5f), la forme parabolique étant cependant moins prononcée pour la variable "sulfates dans les précipitations". En outre, la classe 5 de cette dernière présente sur l'axe 1 une inversion par rapport à la classe 4 indiquant que les plus fortes teneurs de sulfates dans les lacs sont plutôt associées aux valeurs intermédiaires-élevées (classe 4) des sulfates dans les précipitations.

La forte relation observée des teneurs en sulfates dans les lacs avec celles observées dans les précipitations ainsi qu'avec l'éloignement des sources majeures de pollution traduit l'importance du transport atmosphérique à longue distance des composés de soufre dans la composition chimique des eaux lacustres de la région d'étude.

Sulfates - hauteur de précipitation

L'examen des points représentant les classes de précipitation dans le plan 1-2 (Figure 5g) montre la présence d'un lien de cette variable avec les sulfates dans les lacs. Cette relation inverse est cependant présente seulement pour les valeurs de précipitation plus grandes que 900 mm/année (limite supérieure de la classe 1). La position centrale sur le plan 1-2 du point représentant la classe 1 indique une absence de caractéristiques générales pour cette classe. Si on exclut du calcul de la corrélation les valeurs de précipitation correspondant à la classe 1, le coefficient de corrélation de Pearson passe de -0,33 à -0,40 et celui de Spearman de -0,32 à -0,47.

Sulfates - altitude du lac

L'examen du parcours obtenu pour l'altitude du lac (Figure 5b) indique la présence d'un lien de cette variable avec les sulfates dans les lacs. On retrouve en effet une grande similitude de parcours entre ces deux variables. Le coefficient de corrélation est d'ailleurs relativement élevé ($r \cong 0,5$). Dans l'interprétation de la signification de ce lien, il faut cependant tenir compte que la variable altitude n'est pas répartie de façon aléatoire sur la région d'étude. En effet, on retrouve en majorité les lacs de basse altitude dans la partie sud-ouest de la région d'étude et ceux de haute altitude dans la partie nord-est qui

inclut la réserve des Laurentides (LACHANCE et GAMACHE, 1985). Ainsi, on ne peut donc pas discerner l'importance relative de la variable altitude sur la concentration de sulfates en milieu lacustre par rapport aux variables "hauteur de la précipitation" et "concentration de sulfates dans les précipitations" qui suivent un gradient spatial similaire.

Autres relations examinées

L'étude des relations entre variables du milieu et variables de qualité de l'eau s'est poursuivie en considérant le pH et ensuite le calcium + magnésium comme variable à expliquer (LACHANCe, 1988). La même démarche que celle développée précédemment a été adoptée pour chacune de ces variables. On identifie les inter-relations en examinant la projection sur le plan factoriel 1-2 des points représentant les classes des variables et en interprétant les proximités entre points. La description des représentations factorielles obtenues pour le pH ainsi que pour le calcium + magnésium sortirait du contexte de cette étude. C'est pourquoi, on ne présente qu'une synthèse des relations considérées.

Synthèse des relations considérées

Au tableau 2, on identifie les variables du milieu reconnues posséder un pouvoir explicatif des quatre variables de qualité de l'eau considérées ici. Cinq variables du milieu sont reconnues posséder un pouvoir explicatif de l'alcalinité, du pH et du Ca + Mg. Ces variables sont:

- la concentration de sulfates dans les précipitations;
- la hauteur de la précipitation annuelle;
- l'altitude du lac;
- la végétation;
- la géologie.

En ce qui concerne les sulfates, quatre variables ont été identifiées:

- l'éloignement des sources majeures de pollution;
- la concentration de sulfates dans les précipitations;
- la hauteur de la précipitation annuelle;
- la végétation.

Parmi l'ensemble des relations identifiées lors de l'analyse précédente, seulement celles qui ont un sens physique ont été retenues. Par exemple, la relation des sulfates avec la géologie ou celle de l'alcalinité avec l'éloignement des sources n'ont pas été retenues pour cette raison.

5 Élaboration d'un modèle de régression

Dans cette section, on montre que ces relations peuvent être utilisées pour construire un modèle statistique de prévision de variables de qualité de l'eau à partir de certaines variables décrivant le territoire. En particulier, on montre comment l'emploi des techniques d'analyse des données, développées précédemment pour décrire les liaisons entre variables, peut servir également à la prévision d'une variable à partir d'un ensemble de variables explicatives.

Pour décrire les liaisons entre deux groupes K et K' de variables, on a effectué l'analyse des correspondances du tableau $C_{LL'}$ qui a été obtenu en croisant le groupe K de variables ($K \in L$) avec le groupe K' de variables ($K' \in L'$). Pour représenter les individus dans l'analyse, on peut adjoindre au tableau $C_{LL'}$, le tableau disjonctif complet $k_{IL'}$ en tableau supplémentaire. On peut montrer que les facteurs $\{F_{C\alpha}(j) \mid j \in L\}$ obtenus de l'analyse du tableau $C_{LL'}$ peuvent s'interpréter comme la moyenne des facteurs $F_{C\alpha}(i)$ associés aux individus supplémentaires.

Le tableau des facteurs $F^J_{C\alpha}$ peut servir à établir une relation permettant de prévoir une variable y à partir d'un ensemble de variables explicatives. La variable y peut être une des variables à expliquer suivantes: alcalinité, sulfates, pH, Ca + Mg. Pour obtenir une relation permettant d'expliquer ou de prévoir la variable y, on fait une régression multiple usuelle sur les facteurs $F_{C\alpha}$. La relation obtenue peut s'exprimer de la façon suivante:

$$y(i) \approx b_0 + \sum_{\alpha \in A'} b_\alpha F_{C\alpha}(i) \qquad (1)$$

où A' désigne l'ensemble des facteurs conservés.

La régression par voisinage (BORDET, 1973; CAZES, 1976) peut également être utilisée pour prédire la valeur de y au point i. Cette méthode consiste à rechercher les s points les plus proches du point i et à estimer la valeur de y par la moyenne des valeurs de y aux s points voisins de y. Alors que la régression usuelle fournit une précision uniforme (le même écart type), la régression par boule donne une précision qui dépend du point i où l'on se place pour prédire y.

Tableau 2: Identification des variables du milieu qui ont un pouvoir explicatif des quatre variables de qualité de l'eau.

	alcalinité	sulfates	pH	Ca + Mg
Éloignement des sources		√		
Concentration de sulfates dans les précipitations	√	√	√	√
Hauteur de la précipitation Annuelle	√	√	√	√
Dépôt atmosphérique de sulfates				
Altitude du lac	√		√	√
Superficie du lac (SLC)				
Superficie du bassin versant (SBV)				
SBV/SLC				
Pente du bassin				
Abondance des dépôts				
Végétation	√	√	√	√
Géologie	√		√	√

√ : Présence d'une relation significative et interprétable.

La relation précédente, obtenue par la régression multiple usuelle, peut être transformée pour permettre d'obtenir la valeur de y pour un individu nouveau s pour lequel seules sont connues les variables explicatives. Il suffit alors de rajouter l'observation s en ligne supplémentaire à $C_{LL'}$ et de calculer les $F_{C\alpha}(s)$:

$$F_{C\alpha}(s) = \frac{1}{\#Q_e \sqrt{v_\alpha}} \sum_{j \in J_e} k(s,j)\, G_{C\alpha}(j) \qquad (2)$$

où

$k(s,j)$ vaut 0 ou 1 selon que l'individu a adopté la modalité j;

J_e est l'ensemble des modalités des variables explicatives;

$\#Q_e$ est le nombre de variables explicatives;

$G_{C\alpha}(j)$ est le tableau des facteurs issus de l'analyse de C_{LL};
v_α est la valeur propre issue de l'analyse de C_{LL}.

La valeur de cette expression peut être insérée dans la relation (1) pour obtenir y(i).

5.1 Prévision des variables de qualité

La régression multiple usuelle est appliquée en utilisant un programme informatique tiré de la bibliothèque ADDAD version 1983. Pour le calcul de la régression multiple, les deux premiers facteurs ont été retenus. Le test de Student sur le coefficient du troisième facteur a montré que celui-ci n'était pas significativement différent de zéro au niveau de 5%. Pour les variables alcalinité et sulfates, la valeur de t calculée pour le coefficient du deuxième facteur est également inférieure à la valeur théorique pour un niveau de signification de 5%. À la rigueur, il faudrait donc ne considérer que le premier facteur dans l'équation de régression pour ces deux variables.

Les valeurs de R^2 obtenues pour chacune des variables de qualité de l'eau (Tableau 3) sont généralement faibles (0,28 à 0,56) indiquant que l'équation de régression explique faiblement mais significativement le comportement de ces variables. Les meilleures prévisions sont obtenues pour les variables sulfates et calcium + magnésium.

Tableau 3: Régression multiple usuelle sur les variables de qualité de l'eau: valeurs des paramètres de l'équation de régression

		variable à expliquer			
Régression sur le tableau	Paramètre	alcalinité	sulfates	Ca + Mg	pH
$F^I_{C\alpha}$	b_0	94,3	102,0	177,0	5,91
	b_1	-89,1	-40,1	-106,0	-0,66
	b_2	-12,5*	-3,92*	-18,2	0,26
	R^2	0,35	0,56	0,53	0,41
$F'^I_{C\alpha}$	b_0	94,3	102,0	177	5,91
	b_1	-57,3	-27,8	-68,9	-0,36
	b_2	7,1*	3,72*	-19,2	0,03
	R^2	0,28	0,55	0,47	0,28

$F^I_{C\alpha}$: tableau des facteurs des observations supplémentaires de l'analyse de C_{LL} (59 colonnes, 8 lignes: alcalinité, Ca + Mg, 9 lignes: sulfates, pH).

$F'^I_{C\alpha}$: tableau des facteurs des observations supplémentaires de l'analyse du tableau C_{LL} réduit (20 colonnes pour les sulfates, 25 colonnes pour les autres).

* non significativement différent de zéro au niveau de 5% (test de Student)

La qualité de la prévision est légèrement plus faible lorsque l'équation de régression est calculée à partir des résultats de l'analyse du tableau comprenant seulement les variables les plus ex-

plicatives. Les résultats, cependant, confirment l'importance du pouvoir explicatif des variables retenues. En ce qui concerne la variable sulfates par exemple, on peut considérer que près de 55% de la variance des sulfates dans les lacs est expliquée par quatre variables du milieu, soit la concentration de sulfates dans les précipitations, la hauteur de la précipitation annuelle, l'éloignement des sources de pollution et le type de végétation présent sur le bassin versant.

5.2 Exemple d'application

Pour illustrer l'intérêt et la facilité d'utilisation du modèle régressif construit précédemment, on effectue une application de ce modèle à un cas type. Soit un lac inconnu s situé à l'intérieur de la région d'étude dont on voudrait estimer l'alcalinité et les sulfates. On connait seulement les caractéristiques suivantes du milieu:

- la concentration de sulfates dans les précipitations (CS): 40 µg/L;
- la longitude et la latitude du lac: 72°48' et 47°13';
- la précipitation annuelle (PR): 950 mm;
- l'altitude du lac (AL): 365 m;
- le type de végétation sur le bassin versant (VG): feuillu dense;
- la géologie du substrat rocheux du bassin versant (GE): monzonite.

À partir des coordonnées de longitude et latitude, on peut déduire l'éloignement par rapport aux sources majeures de pollution (EL): 1790 d'après l'échelle arbitraire utilisée.

En premier lieu, on identifie, à partir des seuils définis pour les classes, la modalité de chacune de ces variables. On a les classes suivantes: CS2, EL3, PR2, AL3, VG4, GE4. On calcule ensuite la coordonnée $F_\alpha(s)$ de ce point supplémentaire s dans le plan factoriel à partir des résultats des analyses factorielles. Les $G_{C\alpha}^J$ obtenus des analyses des tableaux C_{LL} réduits servent à calculer cette coordonnée $F_\alpha(s)$. Les coordonnées $G_{C\alpha}^J$ ainsi que les valeurs propres v_α obtenus de l'analyse des tableaux pour l'alcalinité et les sulfates sont présentés au tableau 4. On n'a besoin que de la coordonnée correspondant à la classe pour laquelle k(s,j) = 1. Les coordonnées $F_1(s)$ et $F_2(s)$ calculées à partir de la formule (2) pour l'alcalinité et les sulfates sont les suivantes:

	axe 1	axe 2
alcalinité	-0,0315	0,2080
sulfates	-0,1760	0,0560

On peut maintenant utiliser la relation (1) pour calculer par la régression multiple usuelle les estimations des variables alcalinité et sulfates. Les coefficients b de cette équation sont obtenus du tableau 3. On a:

$$\text{alcalinité} = 94{,}3 - 57{,}3\ F_1(s) + 7{,}1\ F_2(s)$$
$$\text{sulfates} = 102 - 27{,}8\ F_1(s) + 3{,}7\ F_1(s)$$

Les valeurs approchées pour l'alcalinité et les sulfates sont 97,6 et 107,1 µg/L respectivement. Ces estimations se comparent avantageusement à des valeurs obtenues pour des lacs voisins possédant des caractéristiques du milieu semblables.

On retient donc qu'il est possible de prévoir la qualité de l'eau des lacs d'une région donnée en connaissant quelques caractéristiques du territoire qui sont reconnues posséder un bon pouvoir explicatif de la qualité de l'eau. La technique qui a été appliquée ici est une régression multiple sur les facteurs d'une analyse factorielle. Cette technique est un outil puissant pour effectuer ce type de prévision. Le modèle élaboré fournit encore une faible qualité d'explication pour une bonne prévision. Cependant, la qualité de la prévision pourrait sûrement être augmentée en améliorant la précision de certaines caractéristiques du territoire et en disposant d'autres variables explicatives non disponibles pour cette étude (type de sols, temps de rétention de l'eau, etc).

Tableau 4: Facteurs utilisés pour l'exemple d'application du modèle de régression.

	alcalinité		sulfates	
Paramètre	axe 1	axe 2	axe 1	axe 2
valeur propre	0,160	0,065	0,401	0,115
CS2	0,383	0,184	0,567	-0,064
PR2	-0,461	0,145	-0,530	-0,266
EL3			0,029	0,428
VG4	-0,265	-0,011	-0,512	-0,022
AL3	0,056	-0,283		
GE4	0,336	0,231		

CS2: concentration de sulfates dans les précipitations - classe 2
PR2: précipitation annuelle - classe 2
EL3: éloignement des sources de pollution - classe 3
VG4: feuillu dense
AL3: altitude - classe 3
GE4: monzonite

6 Conclusion

L'emploi combiné de l'analyse factorielle des correspondances et de la classification ascendante hiérarchique s'est avéré un outil puissant pour synthétiser l'information spatiale, découper la région étudiée en cinq zones homogènes et décrire ces zones en termes de caractéristiques globales par rapport à un ensemble de variables physico-chimiques indicatrices du degré de sensibilité ou d'acidification des eaux.

L'imposition d'une contrainte de voisinage à la méthode de classification hiérarchique a permis de redéfinir les zones de qualité en forçant le regroupement de points voisins et d'obtenir ainsi une délimitation géographique objective des cinq zones. En réduisant au minimum la sévérité de a contrainte de contiguïté, les cinq nouvelles zones redéfinies ont perdu quelque peu de l'homogénéité, mais ont conservé les caractéristiques globales des anciennes zones.

Pour mettre en évidence les variables du territoire qui ont un pouvoir explicatif des variables de qualité de l'eau, on a utilisé l'analyse des correspondances multiples qui est en quelque sorte une application particulière de l'analyse factorielle des correspondances. Appliquée à un tableau croisant les modalités des variables physico-chimiques avec les modalités des variables décrivant le territoire, cette méthode permet de faire ressortir les facteurs biophysiques du territoire qui ont un pouvoir explicatif des variables physico- chimiques indicatrices de la sensibilité ou de l'acidification du milieu aquatique. Parmi ces variables explicatives, on a identifié que la concentration de sulfates dans les précipitations, la hauteur de la précipitation annuelle et le type de végétation sur le bassin versant sont des facteurs explicatifs du pH, de l'alcalinité, du calcium + magnésium et des sulfates dans les lacs. Pour les sulfates, s'ajoute comme facteur explicatif l'éloignement des sources majeures de pollution tandis que pour les trois autres variables (pH, alcalinité, calcium + magnésium), les facteurs suivants qui

s'ajoutent sont l'altitude du lac et la géologie du substrat rocheux du bassin versant.

L'analyse des correspondances multiples permet de représenter les observations individuelles comme des observations supplémentaires dans l'analyse des tableaux croisés. Une régression multiple appliquée aux facteurs représentant ces individus permet de construire un modèle régressif qui peut être utilisé pour prédire une variable physico-chimique d'un lac à partir d'un ensemble de caractéristiques biophysiques explicatives. Actuellement, le modèle fournit encore une faible qualité d'explication des variables physico-chimiques, mais peut être amélioré en considérant d'autres variables du territoire non disponibles présentement. Cette approche de prévision de variables de qualité de l'eau en fonction de caractéristiques du territoire est prometteuse et peut facilement être transposée à d'autres territoires plus vastes. Elle revêt donc une grande importance actuellement en raison de la menace que constitue l'acidification des précipitations pour les écosystèmes sensibles de l'est du Canada et en raison de la grande étendue du territoire menacée.

Références

ADDAD (1983). Logiciel d'analyse des données. M.O. Lebeaux, Laboratoire de statistique, Tour 45-55, 2ième étage, 4, Place Jussieu, 75005 Paris.

ALTSHULLER, A.P. et G.A. MCBEAN (1979). *Le transport des polluants atmosphériques sur de grandes distances en Amérique du Nord: un examen préliminaire*. Préparé par le Groupe consultatif de recherche Etats-Unis - Canada sur le transport des polluants atmosphériques sur de grandes distances. Environnement Canada, 46 p.

ALTSHULLER, A.P. et G.A. MCBEAN (1980). *Deuxième rapport du groupe consultatif de recherche Etats-Unis - Canada sur le transport à grande distance de polluants atmosphériques*. Environnement Canada, 39 p.

BENZÉCRI, J.P. (1977). Sur l'analyse des tableaux binaires associés à une correspondance multiple. *Cahiers de l'Analyse des Données*, 2: 55-71.

BENZÉCRI, J.P. et Coll. (1976). *L'analyse des données. Tome II: l'analyse des correspondances*. Dunod, Paris, 2e édition, 619 p.

BOBÉE, B. M. LACHANCE, J. HAEMMERLI, A. TESSIER, J.Y. CHARETTE et J. KRAMER (1983). *Évaluation de la sensibilité à l'acidification des lacs du sud du Québec et incidences sur le réseau d'acquisition de données*. INRS-Eau, rapport scientifique no 157, 198 p., 4 annexes, 1 appendice.

BOBÉE, B., Y. GRIMARD, M. LACHANCE et A. TESSIER. (1982). *Nature et étendue de l'acidification des lacs du Québec*. INRS-Eau, rapport scientifique no 140, Environnement Québec, publication No PA-1, Envirodoq 351, 243 p., 3 annexes.

BORDET, J.P. (1973). *Étude des données géophysiques - Modélisations statistiques par régression factorielle*. Thèse de doctorat, Université de Paris VI, 242 p.

CAZES, P. (1976). Régression par boule et par l'analyse des correspondances. *Revue de statistique appliquée*, 24: 5-22.

CAZES, P. (1980). L'analyse de certains tableaux rectangulaires décomposés en blocs: Généralisation des propriétés rencontrés dans l'étude des correspondances multiples. II Questionnaires: variantes de codage et nouveaux calculs de contributions. *Cahiers de l'Analyse des Données*, 5: 387-403.

CAZES, P. (1983). Analyse des correspondances multiples: application à l'étude des questionnaires. *Bulletin no. 12, de l'A.D.D.A.D.* (Tour 45-55-E2, 4, Place Jussieu, 75230 Paris, Cedex 05).

COWELL, D.W., A.E. LUCAS et C.D.A. RUBEC (1981). *The development of an ecological sensitivity rating for acid precipitation impact assessment*. Environment Canada, Lands Directorate, working paper no 10, 42 p.

ÉTATS-UNIS - CANADA (1983). *Impact assessment work group 1, Memorandum of intent on transboundary air pollution*. Final report January 1983, 626 p.

GALLEGO, F.J. (1982). Codage flou en analyse des correspondances. *Cahiers de l'Analyse des Données*, 7: 413-430.

GRELET, Y. et M.O. LEBEAUX (1980). Lecture commentée de sorties de programmes en analyse de données. *Bullletin no. 11 de l'A.D.D.A.D.* (Tour 45-55-E2, 4, Place Jussieu, 75230 Paris cedex 05).

GRIMARD, Y. (1984). *Réseau d'échantillonnage des précipitations du Québec: sommaire des données de la qualité des eaux de précipitations 1981- 1982- 1983*. Ministère de l'Environnement du Québec, service de la qualité des eaux, rapport PA-11, Envirodoq 840595, 163 p.

HARVEY, H.H., P.T. DILLON, J.R. KRAMER, R.C. PIERCE et D.M. WHELPDALE. (1981). *Acidification in the canadian environment: scientific criteria for assessing the effects of acidic deposition on aquatic ecosystems*. Associate committee on Scientific criteria for environmental quality, CNRC no 18475, Ottawa, 369 p.

HENDREY, G.R., J.N. GALLOWAY, S.A. NORTON, C.L. SCHOFIELD, P.W. SCHEFFER, D.A. BURNS. (1980). *Geological and hydrochemichal sensitivity of the eastern United States to acid precipitation*. Prepared for the USEPA Ecological effects division, Corvallis Environnement Research laboratory under EPA-DOE interagency agreement, Dept. of Energy and Environment, Brookhaven National Laboratory.

HUNSAKER, C.T., J.L. MALANCHUK, R.J. OLSON, S.W. CHRISTENSEN, ET R.S. TURNER. (1986). Adirondack headwater lake chemistry relationships with watershed characteristics. *Water Air Soil Pollut.*, 31: 79-88.

JAMBU, M. (1976). Sur l'interprétation mutuelle d'une classification hiérarchique et d'une analyse des correspondances. *Revue de Statistique Appliquée*, 24: 45-73.

JAMBU, M. et M.O. LEBEAUX (1983). *Cluster analysis and data analysis*. Elsevier - North-Holland, Amsterdam, 898 p.

KLOPATEK, J.M., W.F. HARRIS et R.J. OLSON (1980). *A regional ecological assessment approach to atmospheric deposition: effects on soil systems*. Dans Shriner, D.S. et al. (éditeurs). Atmospheric Sulphur Deposition: Environment Impact and Health Effects, Ann Arbor Science Publishers Inc. Ann Arbor (Michigan), pp. 539-553.

LACHANCE, M. (1988). *Élaboration d'une méthodologie d'évaluation de la qualité de l'eau: application au problème de l'acidification des eaux lacustres du Québec*. Ecole Nationale Supérieure des Mines de Paris, thèse de doctorat.

LACHANCE, M. et D. GAMACHE (1985). *Sensibilité à l'acidification des lacs du bouclier canadien: analyse et interprétation de données géologiques et biophysiques de 158 bassins versants du sud du Québec*. Environnement Canada, Direction générale des terres, document de travail no 41, 102 p. 3 annexes.

LEBART, L. (1978). Programme d'aggrégation avec contraintes. *Cahiers de l'Analyse des Données*, 3: 275-287.

LEBART, L., A. MORINEAU et J.P. FÉNÉLON (1979). *Traitement des données statistiques - Méthodes et programmes*. *Dunod*, Paris, 510 p.

LOUCKS, O.L., G.E. GLASS, J.A. SORENSEN, B.W. LIUKKONEN, J. ALLERT ET G. RAPP Jr (1986). Role of precipitation chemistry versus other watershed properties in Wisconsin lake acidification. *Water Air Soil Pollut.*, 31: 67-77.

McFEE, W.W. (1980). *Sensitivity of soil to acid precipitation, USEPA-600/3-80-013*. Corvallis Environmental Research Laboratory. Corvallis (Oregon), 186 p.

NORTON, S.A. (1980). *Geologic factors controlling the sensitivity of aquatic ecosystems of acidic precipitation*. Department of Geological Sciences, University of Maine at Orono, Internal report.

PARADIS, D. (1985). *Variabilité spatio-temporelle des précipitations acides sur le Québec méridional (1982-1983)*. Mémoire de maîtrise, Université de Sherbrooke, Québec, 221 p.

RAPP, G. Jr, J.D. ALLERT, B.W. LIUKKONEN, J.A. ILSE, O.L. LOUCKS et G.E.

GLASS (1985). Acid deposition and watershed characteristics in relation to lake chemistry in Northeastern Minnesota. *Environ. International*, 11: 425-440.

ROBERTS, D.J. , T. LINDELL et H. KVARNAS (1982). Environmental factors governing regional lake water quality differences. *Statens naturvordsverk SNVPM* 1621, Avdelning/ Byra.

ROBITAILLE, G. (1980). *Acid precipitation and vegetation*. Communication présentée au 73e congrès annuel de l'Association pour le Contrôle de la Pollution de l'air, juin 22-27, 1980. Montréal (Québec), 71p.

SHILTS, W.W. (1981). *Sensibilité de la roche en place aux précipitations acides: modifications dues aux phénomènes glaciaires*. Commission géologique du Canada, Etude 81-14, 7 p., 3 cartes.

WANG, C. et D. R. COOTE. (1980). *Sensitivity classification of agricultural land to long term acid precipitation in eastern Canada*, Unpublished manuscript and map. 16 p.

L'ANALYSE EXPLORATOIRE DES DONNÉES: DE L'ESPACE MULTIDIMENSIONNEL À L'ESPACE GÉOGRAPHIQUE

Guy Lemay,
Département de géographie, Université du Québec à Montréal

1 De la perception des espaces géographique et mathématique

Dans le processus d'évolution de la discipline géographique, on admet de façon générale que depuis une quarantaine d'années la perception des géographes vis-à-vis l'espace a changé radicalement. En effet, depuis la fin de la deuxième guerre mondiale les géographes ont cessé de considérer l'espace géographique comme absolu et commencé à tenir compte peu à peu des mesures relatives de l'espace telles que le temps, les coûts, etc... Parallèlement à cette évolution, le développement des techniques statistiques qui était amorcé depuis le début du siècle a connu un essor remarquable avec l'arrivée des ordinateurs puissants. On pouvait vers la fin des années 50 exécuter et vérifier les premières analyses multivariées complexes qui avaient été conçues plus tôt mais qui jusque là n'avaient pas été expérimentées à cause des moyens de calcul trop limités.

Cette conjoncture simultanée de l'évolution de la discipline géographique et de la statistique a favorisé un essor important vers la compréhension des structures et des processus spatiaux. Si la question fondamentale du géographe est: "Comment les distributions spatiales se structurent-elles et pourquoi sont-elles structurées?", on ne peut toutefois prétendre répondre à cette question sans considérer les autres sciences.

Les géographes, comme les autres scientifiques, utilisent les sciences formelles pour faciliter l'explication des expériences. Comme le souligne ABLER, ADAMS et GOULD (1971) en travaillant à la frontière entre les construits et les concepts, la géographie incorpore dans ses explications des idées développées par les sciences formelles. Ni les géographes ni les autres spécialistes des sciences factuelles ne peuvent opérer sans recourir aux concepts produits pour les mathématiques et la logique. La géographie a été longtemps logique (déductive) et jamais totalement mathématique, mais la récente emphase de l'utilisation des mathématiques signale la reconnaissance qu'un usage plus intensif de cette forme particulière de la science est un prérequis à un progrès plus avancé dans la connaissance de la théorie géographique. Parce que les structures mathématiques et logiques peuvent être évaluées sans référence aux construits auxquels elles sont appliquées, elles accroissent considérablement notre pouvoir à manipuler les construits. On peut ainsi vérifier la validité de nos manipulations sans être distrait par le contenu (DAVIS 1973).

2 Les difficultés géographiques: un poids deux mesures!

Si les progrès significatifs des sciences factuelles reposent sur le recours intensif à l'utilisation des sciences formelles, on constate que la géographie est au prise avec des difficultés spécifiques. Nous pouvons identifier trois difficultés majeures qui somme toute peuvent l'être pour l'ensemble des sciences sociales. Premièrement, un des facteurs fondamentaux qui contribue à empêcher le progrès de la géographie est la difficulté à formuler les problèmes ou à se préoccuper de problèmes trop généraux. Il importe donc de développer des théories qui puissent contribuer à jeter les bases d'une science spatiale plus structu-

École d'été sur l'analyse des données, juin 1988, UQAM, SCAD-CIRADE.

rée. Les constructions théoriques se développent actuellement rapidement dans les sciences sociales. L'effort de théorisation revêt un caractère imminemment important car il est impossible de produire une "**bonne**" science sans théorie. Le deuxième problème concerne la mesure. Les techniques nécessaires à la mesure des observations et des faits ne sont pas très développées. Des progrès ont été marqués, mais l'observation des faits géographiques et sociaux repose encore sur des techniques peu évoluées. Des efforts devraient être apportés pour mettre au point une théorie de la mesure ou une forme de "**design**" de la mesure. A cet égard les mathématiciens ne pourront rester étrangers à cette nécessité. Finalement, le troisième problème concerne la surabondance des données. Les données sont de plus en plus abondantes, les traitements et les manipulations sont grandement facilités par les progiciels statistiques. Il importe dorénavant d'assurer une bonne sélection des faits qui peuvent être tirés à volonté des banques de données. Dans la prochaine décennie, le défit sera de gérer l'information quantitative et d'en assurer la pertinence et la fiabilité.

3 L'espace géographique multivarié

La plupart des problèmes en géographie reposent sur des forces qui interréagissent de façon complexe et qui ne peuvent être étudiées et isolées individuellement. Étant souvent dans l'impossibilité de décider de la variable la plus significative, on doit examiner toutes les facettes possibles et extraire *a posteriori* les facteurs majeurs (DAVIS 1973). L'objectif principal est de découvrir les dimensions fondamentales des structures spatiales qui résultent des liens qui unissent les distributions.

Les méthodes multivariées nous permettent de considérer les changements simultanément pour diverses propriétés des espaces étudiés. Les exemples de données appropriées à l'analyse multivariée abondent en géographie: que ce soit dans le domaine de la géographie physique dont les variables peuvent être des phénomènes géologiques ou géomorphologiques, ou que ce soit dans le domaine de la géographie humaine avec des variables socio-économiques et démographiques. Ces méthodes sont extrêmement puissantes et efficaces car avec ces outils, les chercheurs manipulent plus de variables qu'ils ne pourraient en appréhender par eux-mêmes. Elles sont par contre compliquées autant au niveau de leur structure théorique qu'à celui de leur application.

Pour plusieurs de ces procédures, la théorie statistique et les tests n'ont été élaborés que dans le cadre restreint d'une série d'hypothèses. La nature et le comportement de ces tests, dans un contexte moins restrictif (problème du monde réel), ne sont pas prévisibles. En fait, certaines des procédures n'ont aucune base statistique théorique et des tests de signification doivent être développés.

Dans le cadre d'une approche statistique en géographie, le choix des éléments à traiter prend une dimension particulière hautement primordiale et devient une tâche bien délicate. Cette opération est beaucoup moins importante pour les travaux dits qualitatifs puisque la donnée sert d'appui au raisonnement intuitif et à l'observation, tandis que dans le cas d'une démarche systémique et mathématique, la donnée est appelée à subir des transformations, à devenir un élément de génération de résultats intermédiaires. En fait, une certaine potentialité, un devenir, sont exigés de la donnée; elle ne doit surtout pas être sélectionnée uniquement pour ce qu'elle représente (pour sa nature apparente) mais aussi pour ce qu'elle est susceptible de révéler (pour sa nature fondamentale du point de vue du géographe). L'habilité à percevoir l'importance du chiffre est directement proportionnelle à l'expérience professionnelle du géographe, à la justesse de sa perception et à sa capacité d'analyse formelle et mathématique.

La spatialité des événements confère une dimension inhérente à leur identité qui doit être prise en considération pour parfaire la compré-

hension et l'explication. À la limite, tous les faits sont spatialisés; cependant selon le genre d'incident, l'espace y joue un rôle plus ou moins important. En admettant cet énoncé, les relations entre les phénomènes et les variations spatiales forment un système intégré qui entraîne l'étude des structures et de leur étendue sur l'espace. Le système des rapports entre les structures et l'espace est composé de la multiplicité des phénomènes hétérogènes variant dans l'espace. Les faits sont non seulement inter-reliés dans un endroit, mais aussi entre plusieurs endroits.

4 La démarche méthodologique intégrée

L'information doit être aisément accessible et mise sous forme de tableaux pour être traitée. Puisque les données se rapportent à des endroits, des lieux, nous l'appelons la matrice d'information spatiale. Simple tableau à double entrées, dont les colonnes accueillent les variables ou attributs spatiaux, tandis que les rangées reçoivent les unités d'observation. Les unités d'observation sont des entités spatiales que ce soit des villes, des divisions de recensement ou des Municipalités Régionales de Comté (M.R.C.), etc. C'est à ce niveau qu'intervient la notion d'échelle qui aura une incidence directe sur la représentation cartographique.

4.1 La matrice d'information spatiale: un atlas potentiel

Chacune des colonnes de la matrice d'information spatiale représente en définitive la distribution d'un phénomène. Ainsi, il est possible de cartographier chacune des variables et d'obtenir autant de cartes qu'il y a de variables prises en compte dans l'analyse. En somme, la matrice d'information spatiale contient potentiellement les structures fondamentales de l'organisation spatiale qu'il faut mettre en évidence avec les techniques d'analyse univariée, bivariée et multivariée. (Voir figure I en annexe)

Toutefois, avant de quitter l'espace géographique pour rejoindre l'espace mathématique, il s'avère nécessaire à l'occasion d'appliquer des mesures de transformation sur les données brutes pour y effectuer des tests d'hypothèse. Cependant, dans le domaine des distributions spatiales, la décision d'accepter ou de rejeter l'hypothèse est hautement subjective, en dépit du fait que cette subjectivité puisse être marquée par un langage et des techniques statistiques impeccablement objectives (GOULD 1975).

La standardisation ou si l'on préfère la transformation en valeurs centrées réduites **Z**, ramène les distributions sur une même échelle et permet une comparaison de plusieurs variables selon une base commune.

Cette première étape de l'analyse revêt un caractère imminemment important puisque les variables de départ renferment les structures latentes de l'organisation spatiale et qu'on ne saurait découvrir des rapports si les éléments essentiels à leur compréhension n'ont pas été introduits au départ.

4.2 La réduction structurelle

Le rôle de la statistique multivariée est d'identifier l'interdépendance entre des événements spatialement définis et de décrire les complexes élémentaires qui sous-tendent les relations préférentielles de certaines distributions (COSINSCHI-MEUNIER 1983). Les méthodes d'analyses factorielles regroupent un ensemble de techniques particulières qui répondent chacune à des contraintes statistiques précises. Les modèles les plus utilisés dans le monde anglo-saxon sont surtout l'analyse en composantes principales et l'analyse factorielle tandis qu'en France l'analyse des correspondances est plus employée.

Sans entrer dans le détail d'une comparaison technique entre ces méthodes, voici les différences majeures entre elles. La figure I illustre schématiquement l'essentiel. Le point de départ dans l'analyse en composantes principales et l'analyse factorielle est la matrice des corrélations. La différence entre ces deux modèles est

que l'analyse en composantes principales assume toute la variation des variables alors que l'analyse factorielle assume qu'une partie seulement de la variation contenue dans les variables. La première méthode consiste surtout à réduire ou à résumer un ensemble de variables, c'est une approche qu'on peut qualifier de déterministe tandis que la seconde a surtout un rôle de test d'hypothèse et on peut la qualifier d'expérimentale (GODDART et KIRBY, 1976).

En ce qui concerne l'analyse des correspondances, c'est le tableau de contingence qui est à la base de l'analyse et la tâche première est de comparer les profils internes de l'ensemble des observations et de l'ensemble des variables. Ainsi plus une variable (ou une observation, selon qu'on traite les colonnes ou les rangées) a un profil différent, plus elle aura tendance à se détacher des autres et à jouer un rôle important.

À la sortie du processus de diagonalisation, nous obtenons les matrices des saturations (loadings) qui comprennent le poids des facteurs ou composantes pour chacune des variables. Puisqu'on traite des variables, la dimension spatiale, c'est-à-dire celle des observations, disparaît momentanément à ce stade. Les facteurs ainsi obtenus renferment le maximum de la variance et ils sont indépendants entre eux.

Une petite particularité de l'analyse des correspondances concerne le premier facteur dit trivial qui a une valeur propre de 1. On n'en tient pas compte dans l'analyse et on obtient un nombre maximal de m-1 facteurs.

Pour obtenir une meilleure explication des facteurs dans les analyses factorielles, on a recours à la rotation qui permet de passer d'un système de facteurs ne possédant pas une structure simple à un autre système de facteurs où chaque variable s'identifie avec seulement un ou un minimum de facteurs (HARMAN 1970).

On peut affirmer qu'à ce niveau de l'analyse l'espace géographique est quelque peu latent, que l'espace multidimensionnel statistique est prédominent selon le nombre de facteurs retenus. L'orthogonalité, c'est-à-dire l'indépendance entre les facteurs est à notre avis un des apports importants des méthodes quantitatives à la géographie. Cependant, on a souvent reproché que les facteurs issus des analyses faisaient ressortir que des évidences. À ce sujet voici quelques précisions. D'une part, on ne saurait obtenir plus d'information que celle qui a été introduite au départ. La pauvreté des variables génère des résultats décevants surtout si la redondance entre elles est importante. De plus, rien n'apparaît plus évident que lorsqu'on a sous les yeux les résultats. On a souvent remarqué une incapacité à faire ressortir les informations les plus pertinentes d'un ensemble de variables sans recourir aux analyses multivariées. Si les facteurs qui en sortaient prétendaient être trop succincts, voire même élémentaires, ils jetaient au moins les fondements d'une organisation des données.

D'autre part, comme nous le soulignions auparavant, le peu d'assise théorique de la géographie n'est-elle pas la source d'une incapacité à définir et à saisir les structures et les processus d'organisation structuraux de l'espace? N'y a-t-il pas un réductivisme naïf de la part des géographes vis-à-vis leur appréhension de l'espace..!

4.3 L'analyse typologique

Suite à la définition des facteurs, l'intérêt du géographe est de retrouver les poids des facteurs sur ses unités spatiales caractérisées par les variables de départ. Le calcul des poids locaux factoriels (factor scores) permet la cartographie de chaque facteur. Ainsi les cartes représentent un ensemble de variables dont l'aspect synthétique est beaucoup plus important que les variables cartographiées une à une. Toutefois, la lecture est plus difficile et l'interprétation oblige à retourner à la matrice des saturations et aux données originales.

Dans le cas de la représentation visuelle des facteurs, on note que l'analyse des correspondances offre l'avantage de représenter dans la même unité de mesure, l'espace des variables et l'espace des observations. Il est plus aisé ainsi de délimiter les associations explicatives entre les deux ensembles. Pour les analyses en composantes principales et factorielles, les unités des saturations et des "loadings" étant différentes, il est plus difficile d'effectuer une représentation simultanée dans un même système de coordonnées.

La classification automatique, faite bien souvent de manière implicite, permet à l'homme de synthétiser l'information et, par là-même, d'avoir une vision globale de son environnement (CHANDON et PINSON, 1981).

Le recours à la taxonomie, en géographie, décompose un ensemble d'unités spatiales décrit par un ensemble de caractéristiques (variables ou facteurs) en un certain nombre de sous-groupes homogènes. Le fait d'établir des classifications hiérarchiques en aval des analyses factorielles repose sur des méthodes heuristiques de classification. En général, on peut faire entrer dans cette catégorie la plupart des algorithmes de classification actuelle. Si à la limite un excellent processus heuristique peut fournir une solution optimale, il est très difficile de déterminer si elle l'est vraiment ou si elle ne fait qu'y tendre, car cette question demeure beaucoup plus philosophique que mathématique (SCOTT 1971).

Méthodologiquement, nous croyons qu'il est plus adéquat d'appliquer d'abord une classification hiérarchique qui donne toute liberté de couper à l'étape d'aggrégation la plus satisfaisante (géographiquement), pour ensuite vérifier l'homogénéité des groupes par l'analyse discriminante. L'expérience que nous avons dans l'application de ce procédé démontre que 80% des classifications testées à l'aide de l'analyse discriminante se stabilisent en dedans de deux itérations, contrairement à des classifications basées sur des a priori qui exigent six itérations et plus.

L'avantage d'effectuer l'analyse discriminante sur une classification hiérarchique aide grandement à vérifier la pertinence de la typologie et à faire ressortir les cas particuliers dans les affinités de groupements. On déduira aisément qu'il est plus avantageux d'effectuer ces analyses sur des résultats factoriels qui sont en quelque sorte des groupes de variables indépendants (mathématiquement) l'un de l'autre que sur des variables originales parfois influencées par l'autocorrélation.

4.4 La représentation visuelle

Si dans la section de la classification automatique, la question de la subjectivité semble absente, elle transparaît cependant dans le choix des critères de sélection et plus encore, à l'étape de la cartographie.

En effet, le géographe dispose de tous les groupements possibles, selon le critère retenu et par conséquent il existe une typologie directement cartographiable. Le problème le plus pressant concerne le manque de solution analytique quant au nombre de groupes à retenir.

5 Au delà des espaces géographiques et mathématiques

La masse sans cesse croissante d'information spatialisée et la complexité ramifiante des champs d'étude exigent un dispositif opérationnel de traitement et de transformation des données suffisamment souple et rapidement accessible. Les progiciels actuels se rapportent davantage de la notion de banque de programmes qu'à celle de système structuré, il se tisse, derrière cette démarche, une toile de fond logique qui évite de se disperser parmi les techniques tout en laissant une marge de manœuvre. Il revient au chercheur de décider lequel parmi tous ces outils est le plus conforme à la nature du problème étudié et lesquels, parmi les différents résultats obtenus des

divers algorithmes, sont les plus représentatifs de la situation géographique analysée.

On se rend compte qu'il ne suffit plus de mettre au point des techniques d'analyse mais qu'il faut exploiter avec la plus grande intelligence possible le type d'information recueillie. Malgré les outils de plus en plus perfectionnés, les géographes demeurent prisonniers d'une information superficielle bien distante de la réalité sociale sans fondement théorique.

Les progrès de la statistique nous ont orientés dans un cheminement analytique assez rigide mais, l'important n'est-il pas comme le prétend TUKEY (1977) de "COMPRENDRE CE QUE L'ON PEUT FAIRE AVANT D'APPRENDRE A MESURER COMMENT EXACTEMENT ON PENSE L'AVOIR FAIT".

Références

ABLER, R., ADAMS, J.S., GOULD, P. (1971) *Spatial organization: The Geographer's view of the World.* Prentice-Hall, 587 p.

AFIFI, A.A., CLARK, V. (1984) *Computer-aided multivariate analysis, lifetime learning* Publications Belmont, California.

CHANDON, J.L., PINSON S. (1981) *Analyse typologique. Théories et applications.* Masson, Paris.

COSINSCHI-MEUNIER, M., (1983) *Géographie quantitative et théorique.* Cahier de cours, Université de Lausanne.

DAVIS, J.C. (1973) *Statistics and Data Analysis in Geology.* Wiley & Sons.

GODDARD, J., KIRBY, A. (1976) *An Introduction to Factor Analysis.* Catmog n° 7, Geo Abstracts Ltd, University of East Anglia, Norwich, 39 p.

GOULD, P. (1975) Les mathématiques en géographie: révolution théorique ou apparition d'un nouvel outil? *Revue Internationale de Sciences Sociales,* Vol.

HARMAN, H.H. (1970) *Modern Factor Analysis.* The Univ. of Chicago Press.

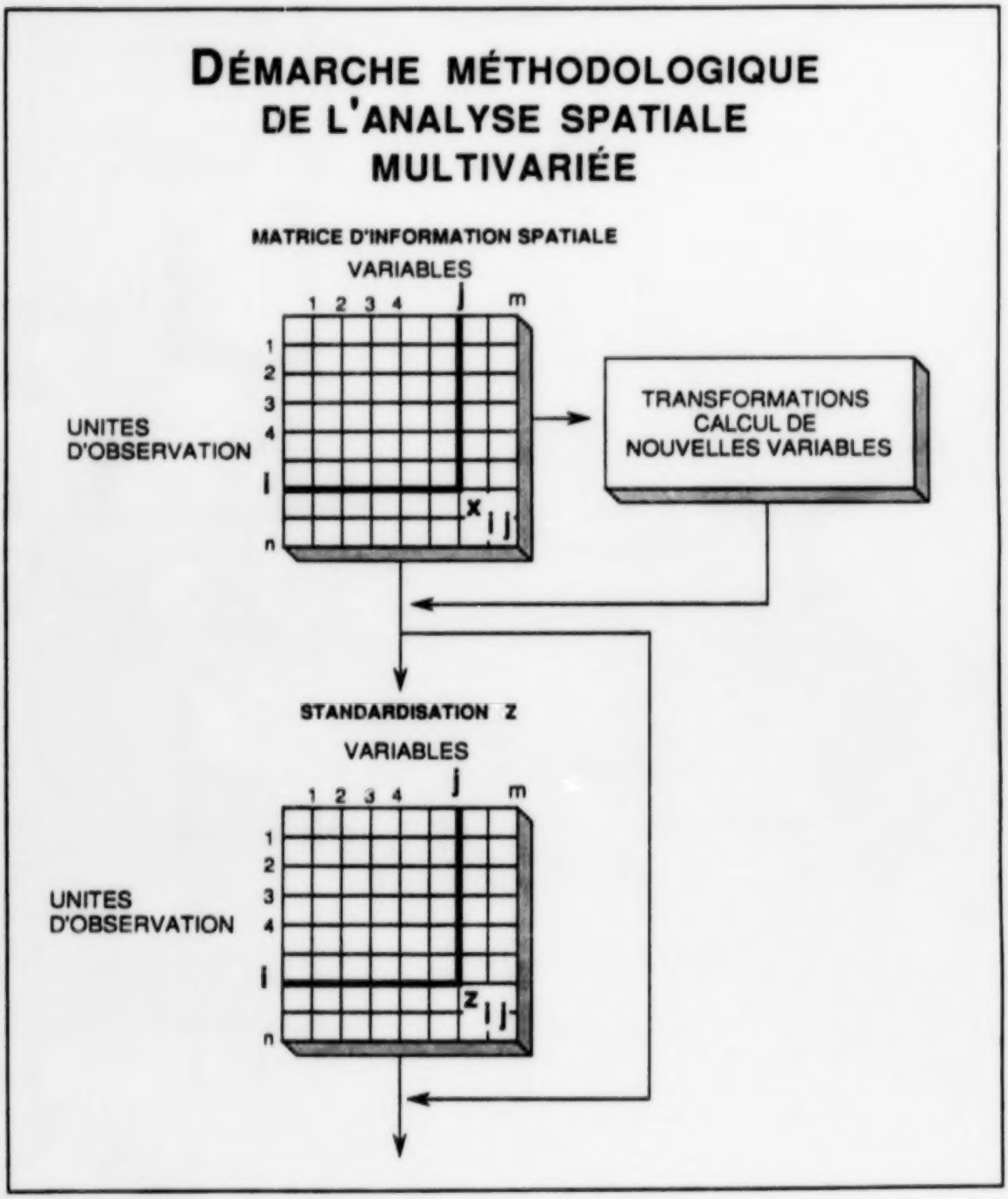
LEBART, L., FÉNELON, J.P. (1971) *Statistique et informatique appliquées.* Paris, Dunod.

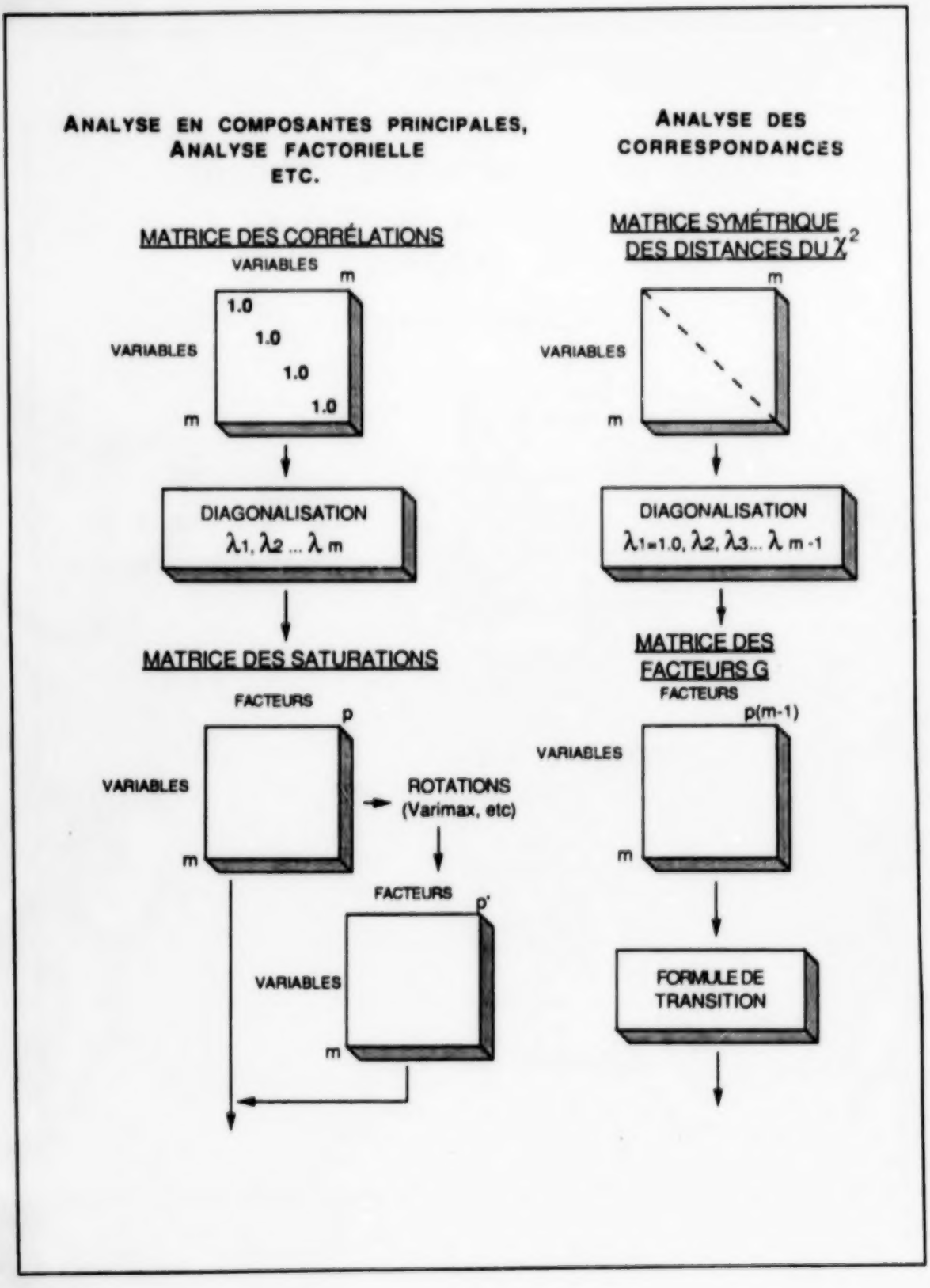
SCOTT, A.J. (1971) *Combinatorial Programming, Spatial Analysis and Planning.* Londres, Nethuen.

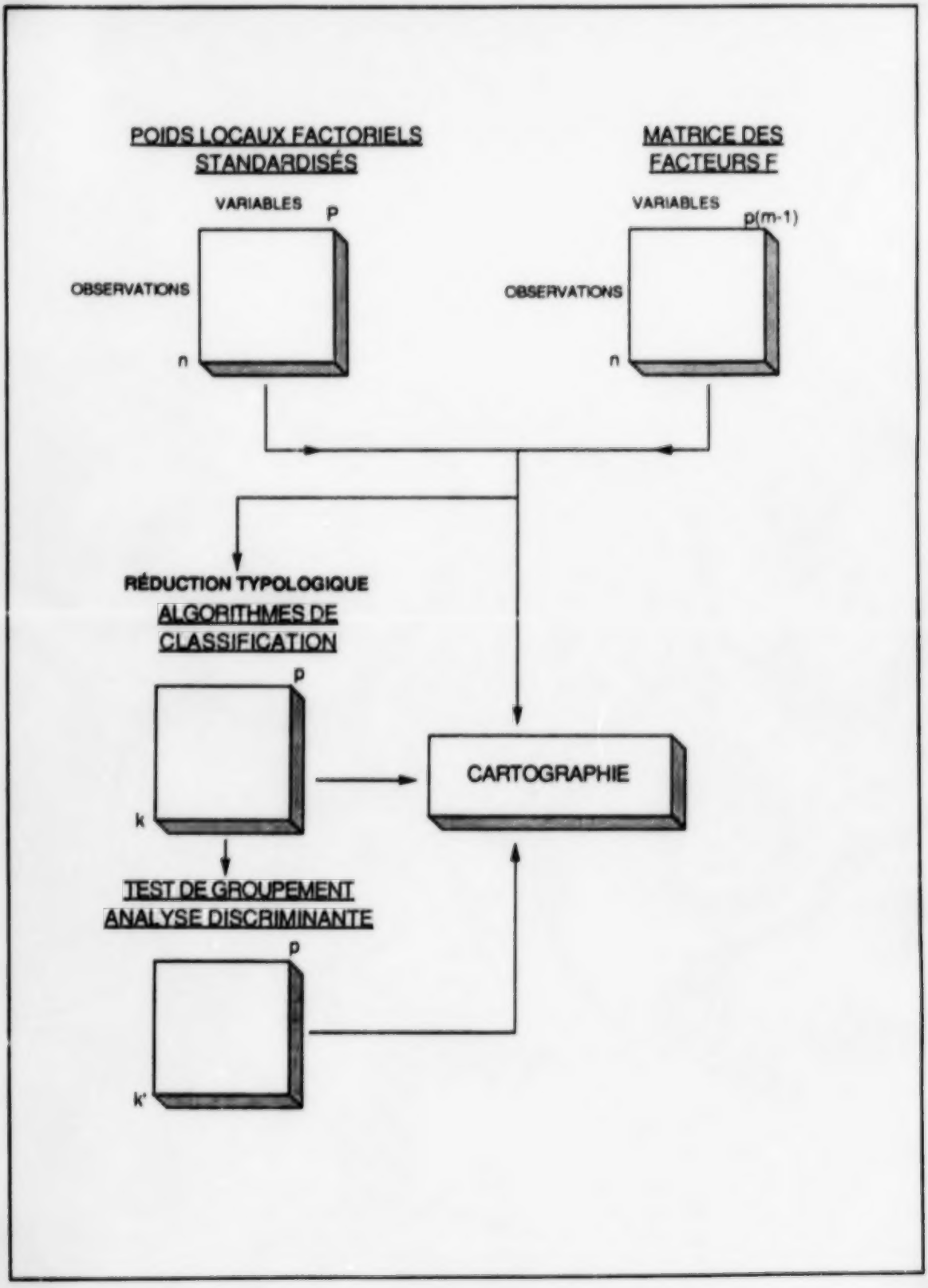
TAYLOR, P.J. (1977) *Quantative Methods in Geography. An Introduction to Spatial Analysis.* Houghton Mifflin Co., Boston, 386 p.

TUKEY, J.W. (1977) *Exploratory date analysis.* Addison-Wesley Publishing Company.

ANNEXE


DÉMARCHE MÉTHODOLOGIQUE
DE L'ANALYSE SPATIALE
MULTIVARIÉE
MATRICE D'INFORMATION SPATIALE
VARIABLES
1 2 3 4 j m
1
2
3
4
i
n
x ij
UNITES
D'OBSERVATION
TRANSFORMATIONS
CALCUL DE
NOUVELLES VARIABLES
STANDARDISATION Z
VARIABLES
1 2 3 4 j m
1
2
3
4
i
n
z ij
UNITES
D'OBSERVATION


ANALYSE EN COMPOSANTES PRINCIPALES, ANALYSE FACTORIELLE ETC.
ANALYSE DES CORRESPONDANCES
MATRICE DES CORRÉLATIONS
VARIABLES
m
1.0
1.0
1.0
1.0
VARIABLES
m
DIAGONALISATION
λ1, λ2 ... λ m
MATRICE SYMÉTRIQUE DES DISTANCES DU χ²
m
VARIABLES
m
DIAGONALISATION
λ1=1.0, λ2, λ3... λ m-1
MATRICE DES SATURATIONS
FACTEURS
p
VARIABLES
m
ROTATIONS
(Varimax, etc)
FACTEURS
p'
VARIABLES
m
MATRICE DES FACTEURS G
FACTEURS
p(m-1)
VARIABLES
m
FORMULE DE TRANSITION


POIDS LOCAUX FACTORIELS STANDARDISÉS
VARIABLES
p
OBSERVATIONS
n
MATRICE DES FACTEURS F
VARIABLES
p(m-1)
OBSERVATIONS
n
RÉDUCTION TYPOLOGIQUE
ALGORITHMES DE CLASSIFICATION
p
k
CARTOGRAPHIE
TEST DE GROUPEMENT
ANALYSE DISCRIMINANTE
p
k'

ÉTUDE DES RÉSIDUELLES DE MODÈLES LOG-LINÉAIRES PAR UNE GÉNÉRALISATION DE L'ANALYSE DES CORRESPONDANCES

Georges Hudon
Département de mathématiques et informatique, Université du Québec à Montréal

1. Introduction

Le but de cette communication est de comparer les résultats de trois méthodes d'analyse différentes appliquées sur un même ensemble de données. Nous choisissons un exemple très simple, un tableau 2×2×2, ce qui permettra une analyse plus facile qu'un gros ensemble de données.

2. Les données

Le tableau 1 nous présente des données relatives à 13,832 meurtres commis aux États-Unis en 1970. Dans cet exemple, nous avons trois variables qualitatives: la race, le sexe et le type de meurtre. Ce qui nous intéresse, c'est d'examiner les relations qui existent entre les trois variables.

Tableau 1: Données relatives aux meurtres commis aux États-Unis en 1970

		Type de meurtre (T)		
Race (R)	Sexe (S)	Fusil (f)	Couteau (c)	Total
Blanche (b)	Homme (h)	3910	808	4718
	Femme (f)	1050	234	1284
Noire (n)	Homme (h)	5218	1385	6603
	Femme (f)	929	298	1227
Total		11107	2725	13832

3. Modèles log-linéaires

Soit x_{ijk} la fréquence observée pour la cellule ijk. Par exemple x_{111} serait la fréquence observée (3910) lorsque Race = Blanche, Sexe = Homme et Type = Fusil. On note les variables par leur première lettre, R = Race, S = Sexe, T = Type et on note aussi $\mu_{ijk} = E[x_{ijk}]$, la fréquence espérée pour la cellule ijk. Le premier modèle que nous examinons est le modèle d'indépendance:

$$\log(\mu_{ijk}) = u + u_{R(i)} + u_{S(j)} + u_{T(k)}$$

École d'été sur l'analyse des données, juin 1988, UQAM, SCAD-CIRADE.

i=1,2 j=1,2 k=1,2, avec les contraintes

$$\Sigma u_{R(i)} = \Sigma u_{S(j)} = \Sigma u_{T(k)} = 0 .$$

Ce type de modèle est en fait un modèle multiplicatif. On veut exprimer la fréquence attendue comme un produit de termes qui ne dépendent que des variables prises individuellement. C'est tout à fait équivalent à supposer que la probabilité conjointe est égale au produit des probabilités marginales, ce qui est la définition même de l'indépendance. Au total, nous avons 4 vrais paramètres: u, $u_{R(1)}$, $u_{S(1)}$, $u_{T(1)}$, les autres paramètres sont obtenus par soustraction. On obtient une estimation de μ_{ijk} par des méthodes bien connues et on le note m_{ijk}. On mesure le degré de divergence entre les fréquences observées et les fréquences attendues par la khideux:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\frac{(x_{ijk} - m_{ijk})^2}{m_{ijk}} .$$

Tableau 2: Fréquences observées et attendues du modèle R,S,T

		Type de meurtre	
Race	Sexe	Fusil	Couteau
Blanche	Homme	3910 (3944.64)	808 (967.78)
	Femme	1050 (874.92)	234 (214.66)
Noire	Homme	5218 (5146.04)	1385 (1262.53)
	Femme	929 (1141.39)	298 (280.04)

On note ce modèle par les variables, soit le modèle R,S,T. Dans notre exemple, nous avons $\chi^2 = 117.02$, ce qui indique que les trois variables qualitatives ne sont pas indépendantes. Le tableau 2 nous donne les fréquences observées avec les m_{ijk} entre parenthèses.

Une fois qu'on a trouvé qu'il y a un lien entre les trois variables qualitatives, il est important de pouvoir spécifier la nature du lien qui existe entre ces variables. Nous regardons maintenant le modèle saturé, c'est-à-dire un modèle qui a autant de paramètres que de cases. On postule

$$\log(\mu_{ijk}) = u + u_{R(i)} + u_{S(j)} + u_{T(k)} + u_{RS(ij)} + u_{RT(ik)} + u_{ST(jk)} + u_{RST(ijk)}$$

avec les contraintes (le + indique une somme sur cet indice)

$$u_{R(+)} = u_{S(+)} = u_{T(+)} = u_{RS(i+)} = u_{RS(+j)} = u_{RT(i+)} = u_{RT(+k)} = u_{ST(j+)} = u_{ST(+k)} = u_{RST(ij+)} = u_{RST(+jk)} = u_{RST(i+k)} = 0$$

On note ce modèle par RST. Il contient l'interaction triple RST et les interactions doubles RS, RT et ST. C'est toujours un modèle multiplicatif, sauf qu'il est saturé, c'est-à-dire que l'on obtient des fréquences attendues égales aux fréquences observées. On a aussi $\chi^2 = 0$. Si on pose $u_{RST(ijk)} = 0$, on obtient un modèle qui ne contient pas l'interaction triple RST mais qui possède les interactions doubles RS, RT et ST. On postule pour ce modèle

$$\log(\mu_{ijk}) = u + u_{R(i)} + u_{S(j)} + u_{T(k)} + u_{RS(ij)} + u_{RT(ik)} + u_{ST(jk)}$$

avec les contraintes appropriées. On note ce modèle RS,RT,ST. Le tableau 3 nous donne les fréquences attendues m_{ijk} découlant de ce modèle.

Pour ce modèle, nous avons $\chi^2 = 1.08$, avec un niveau de signification de 0.2997, ce qui indique que le modèle s'ajuste très bien au données. Les fréquences attendues peuvent s'écrire sous la forme d'un produit, puisque c'est un modèle multiplicatif. Au tableau 4, on présente la décomposition des fréquences attendues.

Tableau 3: Fréquences attendues du modèle RS,RT,ST

		Type de meurtre	
Race	Sexe	Fusil	Couteau
Blanche	Homme	3919.4603	798.53965
	Femme	1040.5397	243.46035
Noire	Homme	5208.5397	1394.4603
	Femme	938.46035	288.53965

Tableau 4: Décomposition des fréquences attendues selon modèle RS,RT,ST

			Effet						
Race	Sexe	Type	Attendu	R	S	T	RS	RT	ST
Blanche	Homme	Fusil	3919.4603=(1057.1)(0.893)(2.066)(1.999)(0.908)(1.071)(1.035)						
		Couteau	798.53965=(1057.1)(0.893)(2.066)(0.500)(0.908)(0.934)(0.966)						
	Femme	Fusil	1040.5397=(1057.1)(0.893)(0.484)(1.999)(1.102)(1.071)(0.966)						
		Couteau	243.46035=(1057.1)(0.893)(0.484)(0.500)(1.102)(0.934)(1.035)						
Noire	Homme	Fusil	3919.4603=(1057.1)(1.120)(2.066)(1.999)(1.102)(0.934)(1.035)						
		Couteau	798.53965=(1057.1)(1.120)(2.066)(0.500)(1.102)(1.071)(0.966)						
	Femme	Fusil	1040.5397=(1057.1)(1.120)(0.484)(1.999)(0.908)(0.934)(0.966)						
		Couteau	243.46035=(1057.1)(1.120)(0.484)(0.500)(0.908)(1.071)(1.035)						

On est maintenant en mesure d'examiner les effets. L'interaction RS nous donne un effectif plus élevé pour une femme blanche et pour un homme noir que les autres, toute proportions

gardées. On voit que l'interaction la plus forte est RS (1.102), la deuxième est RT (1.071) et la plus faible est ST (1.035). Pour vérifier ces résultats, voici les résultats de l'ajustement de plusieurs modèles à ces données.

Modèle	dl	χ^2	niveau de signification
RST	0	0	———
RS,RT,ST	1	1.08	0.2997
RS,RT	2	7.57	0.0227
RS,ST	2	39.92	0.0000
RT,ST	2	78.00	0.0000
R,S,T	4	117.02	0.0000

Le modèle le plus plausible est RS,RT,ST. Par contre le facteur ST ne semble pas être très significatif, puisque le modèle RS,RT donne un niveau de signification de 0.0227. Il faut donc faire très attention si nous rencontrons par hasard une femme blanche avec un fusil ou un homme noir avec un couteau.

4. Analyse des correspondances multiples

Nous allons maintenant examiner les données avec l'analyse des correspondances multiples. On obtient le tableau disjonctif complet suivant:

Étiquette					
Rb	Rn	Sh	Sf	Tf	Tc
3910	0	3910	0	3910	0
808	0	808	0	0	808
1050	0	0	1050	1050	0
234	0	0	234	0	234
0	5218	5218	0	5218	0
0	1385	1385	0	0	1385
0	929	0	929	929	0
0	298	0	298	0	298

Sur ce tableau nous appliquons l'analyse des correspondances. On obtient les valeurs propres suivantes:

No	Inertie	%
1	0.36073	36.073
2	0.33882	33.882
3	0.30046	30.046

Au tableau 5, on retrouve les coordonnées (×100) et les contributions relatives pour les deux premiers axes. La figure 1 nous en fournit une représentation graphique.

Il est clair que l'axe 2 est l'effet ST, les femmes préférant les couteaux. L'axe 1 est la superposition des deux effets RS et RT. Il est problématique de noter que l'effet ST est beaucoup plus faible que les effets RS et RT, mais l'axe 2 représente quand même 33.882 % de l'inertie totale. Il est aussi troublant de noter que si l'axe 1 est formé des deux effets RS et RT, on doit conclure que les femmes sont opposées aux couteaux sur cet axe. On obtient alors une contradiction qui semble assez difficile à résoudre.

Il faut aussi noter que l'analyse des correspondances multiples est une technique bivariée. Cette technique n'est pas en mesure d'évaluer les interactions triples ou quadruples dans un ensemble de données. L'exemple que nous venons d'étudier est clairement un exemple bivarié, car on ne retient pas l'interaction triple. L'analyse des correspondances multiples devrait donc bien fonctionner dans un contexte semblable. Pour examiner plus en détail les interactions doubles, supposons que l'on enlève l'effet ST des données. C'est-à-dire que nous voulons étudier la matrice des fréquences attendues sous le modèle RS, RT. Le tableau 6 donne les fréquences attendues pour ce modèle. Sur ce tableau nous appliquons l'analyse des correspondances multiples. On obtient les valeurs propres suivantes:

No	Inertie	%
1	0.36388	36.388
2	0.33214	33.214
3	0.30398	30.398

Tableau 5: Résultats de l'analyse de correspondances multiples des données observées

	Axe 1			Axe 2	
Étiquette	Coord	Ctrel	Étiquette	Coord	Ctrel
Rb	87.5	30.7	Rb	-8.3	0.3
Rn	-67.1	23.5	Rn	6.3	0.2
Sh	-28.8	6.3	Sh	-26.6	5.7
Sf	129.7	28.2	Sf	120.1	25.8
Tf	17.3	2.2	Tf	-41.2	13.4
Tc	-70.5	9.1	Tc	167.8	54.6

Figure 1: Représentation des variables selon les deux premiers axes

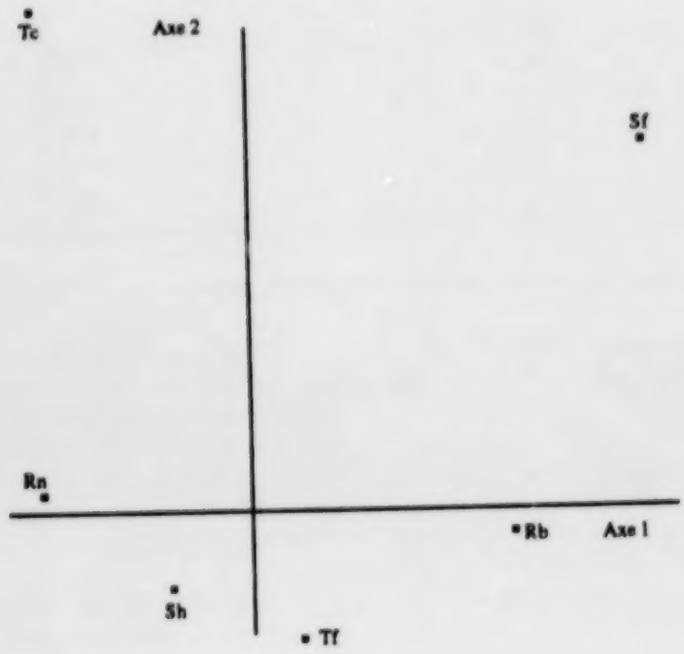


Tableau 6: Fréquences attendues du modèle RS,RT

		Type de meurtre		
Race	Sexe	Fusil	Couteau	Total
Blanche	Homme	3898.9	819.1	4718
	Femme	1061.1	222.9	1284
Noire	Homme	5183.7	1419.3	6603
	Femme	963.3	263.7	1227
	Total	11107	2725	13832

Tableau 7: Résultats de l'analyse de correspondances multiples des données attendues selon le modèles RS,RT

Axe 1			Axe 2		
Étiquette	Coord	Ctrel	Étiquette	Coord	Ctrel
Rb	-83.5	27.7	Rb	-1.7	0.0
Rn	64.0	21.3	Rn	1.3	0.0
Sh	28.5	6.1	Sh	26.9	6.0
Sf	-129.0	27.5	Sf	-121.0	26.9
Tf	-21.6	3.4	Tf	40.5	14.0
Tc	88.1	14.0	Tc	-165.0	53.9

La tableau 7 nous donne les coordonnées (×100) et les contributions relatives pour les deux premiers axes.

Le deuxième axe demeure un axe ST, un effet qui a été enlevé du modèle. Le premier axe demeure aussi un axe RS,RT avec la même contradiction concernant Sf et Tc. Les femmes sont opposées aux couteaux sur l'axe 1, elles sont favorables aux couteaux sur l'axe 2, mais conditionnellement à la race, il n'y a pas de lien entre S et T avec ce tableau de données.

5. Analyse des correspondances multivariées

Escofier (1983) a généralisé l'analyse des correspondances pour examiner la différence entre deux matrices de nombres non-négatifs. Van Der Heijden (1985) a utilisé les tableaux multiples et la généralisation d'Escofier pour décomposer la différence entre deux matrices, chacune distribuée selon un modèle log-linéaire. Dans ses analyses, chaque matrice devait avoir les mêmes marginales.

Ici nous utilisons un cas particulier de la généralisation d'Escofier pour étudier la différence entre la matrice de données et la matrice des fréquences attendues sous l'hypothèse d'indépendance. Dans notre cas, les deux matrices n'ont pas les mêmes marginales. On nomme cette méthode "l'analyse des correspondances multivariées".

Pour le tableau qui nous intéresse, la χ^2 d'indépendance est donnée par la formule suivante:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{(x_{ijk} - x_{i++}x_{+j+}x_{++k}/n^2)^2}{x_{i++}x_{+j+}x_{++k}/n^2}$$

où + indique une somme sur l'indice correspondant et n est le nombre d'observations. On calcule les termes suivants qui interviennent dans le calcul de la χ^2 :

$$\frac{x_{ijk} - x_{i++}x_{+j+}x_{++k}/n^2}{\sqrt{x_{i++}x_{+j+}x_{++k}}}$$

et on forme une matrice **X** avec ces termes, en croisant deux des variables. Par exemple, le tableau multiple RS×T serait

3910	808
1050	234
5218	1385
929	298

et la matrice **X** serait une matrice 4×2 formée des termes appropriés. Le tableau multiple R×ST serait

3910	808	1050	234
5218	1385	929	298

De la matrice **X'X**, nous calculons les valeurs propres. La somme des valeurs propres donne χ^2/n^2. C'est donc une décomposition de la χ^2 d'indépendance. Chaque valeur propre représente l'inertie d'un axe. Dans notre cas particulier nous aurons deux axes. Les coordonnées sont la projection du profil d'une ligne de la matrice **X** sur les vecteurs propres (de norme 1) de la matrice **X'X** ou d'une colonne de la matrice **X** sur les vecteurs propres (de norme 1) de la matrice **XX'**. Il est à remarquer que les valeurs propres non-nulles de **XX'** et **X'X** sont les mêmes. Les contributions aux axes sont le carré des coordonnées pondérées par la masse de la ligne ou de la colonne.

Les valeurs propres que nous calculons dépendent du tableau multiple utilisé. Dans notre exemple, la χ^2 d'indépendance donne 117.02. Voici la décomposition de cette χ^2 selon le tableau multiple utilisé.

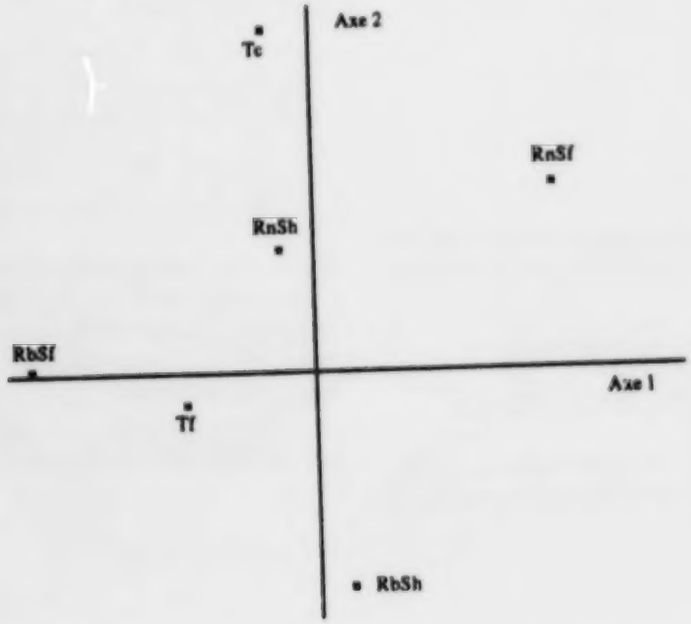
	Axe 1		Axe 2	
Tableau	χ^2	%	χ^2	%
RS×T	77.32	66.1	39.70	33.9
R×ST	112.73	96.3	4.29	3.7
RT×S	81.33	69.5	35.69	30.5

On retrouve au tableau 8 les coordonnées (×100) et les contributions relatives (%) pour les deux premiers axes pour le tableau RS×T. Le tableau R×ST contient toute son inertie sur l'axe 1. Il serait donc préférable d'utiliser le tableau RS×T ou le tableau RT×S. La figure 2 nous fournit une représentation graphique.

Tableau 8: Résultats de l'analyse de correspondances multivariées du tableau RS × T

Axe 1			Axe 2		
Étiquette	Coord	Ctrel	Étiquette	Coord	Ctrel
RbSh	2.23	3.2	RbSh	-7.03	61.1
RbSf	-18.37	47.5	RbSf	0.35	0.0
RnSh	-2.09	3.6	RnSh	3.97	25.5
RnSf	15.77	45.7	RnSf	6.11	13.4
Tf	-8.18	96.1	Tf	-1.19	4.0
Tc	-3.35	3.9	Tc	11.82	96.0

Figure 2: Représentation des variables selon les deux premiers axes



L'axe 1 est formé principalement de l'effet RS (opposition manifeste entre RbSf et RnSf) tandis que l'axe 2 nous révèle l'effet RT (RnSh, RnSf et Tc sont tous positifs). Pour compléter la comparaison, étudions la matrice des fréquences attendues sous le modèle RS,RT, comme nous avons fait pour l'analyse des correspondances multiples. Le tableau 9 nous donne les fréquences attendues.

Tableau 9: Fréquences attendues du modèle RS, RT

		Type de meurtre		
Race	Sexe	Fusil	Couteau	Total
Blanche	Homme	3898.9	819.1	4718
	Femme	1061.1	222.9	1284
Noire	Homme	5183.7	1419.3	6603
	Femme	963.3	263.7	1227
	Total	11107	2725	13832

On fait l'analyse avec le tableau RS×T. Le χ^2 d'indépendance donne 111.79. Voici la décomposition de ce χ^2.

	Axe 1		Axe 2	
Tableau	χ^2	%	χ^2	%
RS×T	74.92	67.0	36.87	33.0

Au tableau 10 on retrouve les coordonnées (×100) et les contributions relatives (%) pour les deux premiers axes pour le tableau RS×T.

On voit encore plus clairement les axes. L'axe 1 représente l'effet RS et l'axe 2 reflète l'effet RT. On voit bien que l'analyse des correspondances multivariées donne un résultat plus satisfaisant que la multiple. Il est donc conseillé d'utiliser la multiple avec beaucoup de prudence.

Tableau 10: Coordonnées (×100) et les contributions relatives (%) pour les deux premiers axes pour le tableau RS×T.

Axe 1			Axe 2		
Étiquette	Coord	Ctrel	Étiquette	Coord	Ctrel
RbSh	3.80	9.5	RbSh	-5.75	44.1
RbSf	-18.02	47.3	RbSf	-6.45	12.3
RnSh	-2.91	7.2	RnSh	4.72	38.8
RnSf	13.77	36.0	RnSf	3.52	4.8
Tf	-7.45	82.4	Tf	-2.42	17.6
Tc	-6.96	17.6	Tc	10.56	82.4

Bibliographie sommaire

BENZÉCRI, J.-P. (1973). *L'Analyse des Données, Tome 2: L'Analyse des Correspondances*. Paris, Dunod.

BISHOP, Y.M.M., FIENBERG, S.E., HOLLAND P. W. (1975). *Discrete Multivariate Analysis*. The M.I.T. Press, Cambridge, Massachusetts.

ESCOFIER, B. (1979). *Traitement Simultané de Variables Qualitatives et Quantitatives en Analyse Factorielle*. Cahiers de l'Analyse des Données, 4, 137-146.

ESCOFIER, B. (1983). *Analyse de la différence entre deux mesures sur le produit de deux mêmes ensembles*. Cahiers de l'Analyse des Données, 3, 325-329.

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London, Academic Press.

GREENACRE, M. J., HASTIE, T. (1987). *The Geometric Interpretation of Correspondence Analysis*. Journal of the American Statistical Association, 82, 437-447.

HUDON, G. (1988). *Multivariate Correspondence Analysis*. Rapport de recherche n°55, département de mathématiques et d'informatique, Université du Québec à Montréal.

LEBART, L., MORINEAU, A., WARWICK, K. (1984). *Multivariate Descriptive Statistical Analysis*. New York, John Wiley.

TENENHAUS, M., YOUNG, F. W. (1985). *An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Data*. Psychometrika, 50, 91-119.

VAN DER HEIJDEN, P. G. M. (1985). *Correspondence Analysis used complimentary to Loglinear Analysis*. Psychometrika, 50, 429-447.

APPLICATIONS DE L'ANALYSE FACTORIELLE À LA LINGUISTIQUE

Marc Bourdeau
École Polytechnique de Montréal

1 Introduction

Les mathématiques ont nourri au lait de la linguistique au moins trois de leurs plus beaux enfants! Disons, de façon plus ... précise, que des considérations de nature linguistique ont inspiré des mathématiciens pour fonder ce qui constitue aujourd'hui trois chapitres importants des mathématiques. A. A. MARKOFF (1913) a eu l'idée d'étudier des événements enchaînés (et non plus indépendants) en considérant les enchaînements des consonnes et des voyelles dans le poème épique "Eugène Onéguine" de Pouchkine, et fonda la théorie des processus stochastiques. Il a laissé son nom, entre autres, aux chaînes qui portent son nom. SHANNON (1954) s'est intéressé quant à lui aux propriétés des codes, ces véhicules nécessaires à toute communication, puisque tout langage en possède nécessairement un. La théorie de l'information vient essentiellement, dans ses fondements, de Shannon. Enfin J.-P. BENZÉCRI, au début des années soixante, alors qu'il était professeur à Rennes, a pensé aux *correspondances* en se penchant sur des données linguistiques. Benzécri a eu indubitablement un rôle prépondérant dans l'établissement de l'Analyse des données, à tout le moins dans l'acception française du terme.

Comme il l'explique lui-même, Benzécri a voulu fonder sur la statistique une linguistique inductive, nous dirions plutôt quantitative aujourd'hui, en la basant sur l'analyse de tableaux de contingences pertinents (Benzécri, 1981). Les observations, et les statistiques qui en dérivent, décrivent les relations entre des termes de nature linguistique. L'objectif est de passer de l'étude d'un corpus à la généralité de la langue. De l'usage à la langue, en somme.

Le terme même de correspondances avait ainsi, à l'origine, une connotation linguistique. Des considérations et liens entrevus sur des tableaux de contingences où, par exemple, les lignes désignent des noms et les colonnes les verbes des contextes où les noms apparaissent dans un certain corpus, il conçut des analyses quantitatives de ces liens, qui expriment les associations entre les deux types d'objets. Les correspondances viennent de là. Quoiqu'il en soit des prédécesseurs de Benzécri, il semble bien que toute la philosophie d'origine des méthodes françaises part de là. On assiste cependant aujourd'hui à une convergence entre les deux grandes écoles de statistiques, la traditionnelle de nature inductive (fishérienne est-on tenté de dire) et la française de nature plus descriptive. Mais la pensée géométrique sous laquelle s'effectue cette synthèse n'aurait probablement pas pu se constituer sans les traverses radicales qu'a su prendre Benzécri.

Le terme de linguistique inductive, utilisé par Benzécri, signifie la possibilité de décrire des propriétés du langage en général par l'observation de corpus forcément limités. Des tentatives de formalisation du phénomène linguistique ont eu lieu très tôt dans l'histoire. Les grammaires sont de telles formalisations. Mais en fait l'observation directe des phénomènes eux-mêmes dans toute leur diversité, qui permet de fait à la linguistique de devenir une science quantitative, n'est possible que depuis peu. Il faut bien noter que la linguistique quantitative ne peut s'exercer

École d'été sur l'analyse des données, juin 1988,
SCAD-CIRADE

que lorsqu'on possède des moyens considérables de traiter l'information. Des tentatives, comme celles de JAKOBSON ET HALLE (l'école du MIT) au début des années cinquante étaient vouées à l'échec, et sont tombées dans l'oubli faute de moyens de calcul et de modèles quantitatifs appropriés (CHERRY *et al.*, 1953).

Le quantitatif en linguistique est ainsi fortement lié aux possibilités de dénombrement et d'ordonnancement que permet l'ordinateur (terme tellement plus exact que *computer*), lequel ne saurait exister, il faut le noter, sans l'exploitation de langages de tous niveaux! En bref, de science humaine la linguistique devient de plus en plus une science quantitative et exacte.

Il est bien évident que nous ne parlerons que d'une infime partie des travaux en linguistique quantitative. Nous rapportons en bibliographie une bonne partie des travaux qu'il est possible de trouver dans les revues françaises d'Analyse des données, en plus bien sûr des références de ce qu'on exposera ici.

Dans ce qui suit, nous avons choisi de restreindre notre panorama à des travaux de linguistique quantitative qui explorent aux niveaux phonologique et phonétique le corpus SANKOFF-CEDERGREN (SC). Cet énorme corpus a été constitué, au début des années soixante-dix sous la direction de deux sociolinguistes —Gillian Sankoff et Henrietta Cedergren—, en vue de l'étude de la variation linguistique à l'intérieur de la communauté francophone de Montréal. DAVID SANKOFF, quant à lui, a eu un rôle important dans le traitement quantitatif du corpus. Celui-ci est basé sur des entrevues de cent vingt individus, sous forme de conversation informelle concernant la vie à Montréal. Sankoff *et al.* (1976, 1978) en font des descriptions assez complètes. Chaque locuteur énonce approximativement huit mille mots, soit environ une heure de "texte", et est caractérisé par des variables socio-économiques: sexe, âge, profession, nombre d'années de scolarité, et un indice socio-linguistique. Un sous-ensemble (la moitié) du corpus d'origine a de nouveau été échantillonné en 1984, et nous parlerons de son exploitation plus loin.

En premier lieu, nous utilisons des variantes d'analyses factorielles pour mettre en évidence des propriétés phonologiques des discours des locuteurs, et nous examinons des analyses discriminantes pour séparer les parlers (formel, familier ou populaire) utilisés. Notre étude repose sur des caractéristiques quantitatives globales définies sur la chaîne phonologique (les phonèmes) des locuteurs, de façon tout à fait différente de celles des études qui sont basées sur des dénombrements et des tableaux de contingences particuliers (*e.g.* ROUSSEAU ET SANKOFF, 1978; MORIN, 1983).

À un autre niveau, celui de l'analyse spectrale, la chaîne phonémique est représentée en termes des fréquences des premières harmoniques du signal sonore des voyelles dans leur contexte d'enchaînement. Nous décrirons sommairement une application très directe de l'analyse factorielle sur des variables fréquentielles des spectres phonémiques. Ce dernier travail est encore dans ses premiers balbutiements: la méthodologie n'est pas parfaitement arrêtée, la cueillette des données est fort longue. C'est l'avenir qui montrera si l'analyse de données y a été utile.

2 Analyse factorielle et phonologie

Beaucoup d'études ont mis en action des "machineries" très élaborées sur, plus ou moins directement, des tableaux de contingences ou des comptages. À l'inverse, nous utiliserons des méthodes usuelles de statistique multidimensionnelle, mais nous définirons des variables de façon beaucoup moins directe, bien qu'elles représentent des propriétés simples des langages.

Un langage apparaît toujours comme un enchaînement dans les choix d'unités. En fait, il y a une sorte de hiérarchie de ces unités, et on parle parfois de la pyramide langagière: en simplifiant quelque peu, des phonèmes constituent des syl-

labes, celles-ci des mots, de mots font des phrases, etc. Nous nous intéressons à la chaîne phonologique (BOURDEAU, 1986, 1987). Dénombrer et comparer sont en fait des processus très primitifs de la pensée, mais néanmoins à la base de toute science quantitative (est-ce qu'il y en a d'autres?). Ainsi, à la considération d'un langage en tant que chaîne d'événements, un esprit quantitatif se demande combien d'unités différentes apparaissent dans un échantillon donné, quelles en sont les proportions relatives, quelles sont les contraintes qu'un langage impose au processus d'enchaînement des unités. On peut penser en ces termes à tous les niveaux de la pyramide. Il ne peut évidemment être question de considérer à la fois toutes les chaînes possibles de noms, par exemple, mais on peut définir des mesures globales des phénomènes qui apparaissent, et espérer que celles-ci fournissent un modèle cohérent de la réalité à décrire.

Dans notre exemple, les unités sont les phonèmes. Les phonèmes sont les lettres de notre écriture orale. Un échantillon des phonèmes émis par un locuteur fournit des estimations des fréquences relatives des phonèmes de la langue utilisée. Si cet échantillon est suffisamment long, ces estimations sont très précises.

On pose: $P=\{p_1, \ldots, p_n\}$, la répartition des phonèmes d'un échantillon de longueur donnée. L'entropie d'une répartition est une mesure de l'écart de cette distribution à la distribution uniforme. On utilise cette mesure également pour quantifier la diversité dans un échantillon d'espèces (par exemple), lorsqu'on traite de leur abondance (YAGLOM et YAGLOM, 1969). Pour tenir compte des contraintes imposées par l'enchaînement des unités phonologiques, on est appelé à définir des entropies d'ordre supérieur. Compte tenu de la précision que nous désirons, et de la taille échantillonnale qui est accessible, nous avons fixé la longueur de nos échantillons à 1200 phonèmes. Ceci nous donne une taille échantillonnale de 388; chacun des sujets du corpus SC fournit environ 3 échantillons. Au total, nous disposons de la plus grande transcription phonologique du français. Il est inutile de rappeler ici les définitions (voir les références). Contentons-nous de dire que nous avons calculé pour ces 388 "sujets" les entropies jusqu'à l'ordre trois. Ces variables, normalisées par l'entropie maximale de chaque échantillon, nous donnent les *redondances* des échantillons jusqu'au même ordre: R_1, R_2, R_3. Ce sont nos premières variables de description des locuteurs.

Ces variables caractérisent globalement les propriétés de sélection des unités dans la chaîne parlée. On ne peut considérer à la fois tous les enchaînements, cela va de soi, mais on peut catégoriser ce continuum (vieille astuce des statisticiens!). Pour décrire certaines propriétés d'enchaînement, on peut séparer les phonèmes en classes naturelles et se restreindre aux enchaînements des classes. Les consonnes et les voyelles fournissent évidemment une première classification des phonèmes. Nous avons donc tiré de chacune des 388 chaînes la matrice stochastique des transitions, T_2 des voyelles et des consonnes. Schématiquement, on écrit:

$$T_2 = \begin{pmatrix} CC & CV \\ VC & VV \end{pmatrix},$$

où CC+CV, somme des proportions des enchaînements de consonnes à consonnes et de consonnes à voyelles, vaut l'unité. Il en est de même pour la deuxième ligne. Donc deux des variables sont statistiquement redondantes, et on n'en a gardé que deux, soit CV et VC. Mais pour tenir compte des liens géométriques dans T_2 qui ne sont pas contenus dans CV et VC, nous avons calculé, pour chacune de ces matrices stochastiques, la valeur propre autre que l'unité, λ, et l'angle θ entre les deux vecteurs propres. On montre que:

$$\lambda = 1 - (CV+VC),$$
$$\theta = \arctan(-VC/CV).$$

Ces deux dernières variables font jouer CV et VC, et comme CVC forme la syllabe "élémentaire", on les appelle les *facteurs syllabiques* de l'échantillon.

Beaucoup de nos analyses reposent sur les sept variables descriptives que nous venons de définir. Ce sont, remarquons-le encore, des caractéristiques globales de la sélection et l'enchaînement des unités: aucune considération sur tel ou tel phonème n'entre en ligne de compte.

Donc, d'une part sept variables de sélection et d'enchaînement des unités. Mais, d'autre part, pour rentrer un peu plus finement dans la phonologie des sujets nous avons calculé aussi les matrices de transition des phonèmes lorsqu'on les partage en six classes cette fois, en se fondant sur les modèles usuels de la phonologie et sur des considérations articulatoires. On a fixé ainsi trois classes de consonnes et trois de voyelles. On dispose ainsi de 388 matrices, notées T_6, et donc de 36 variables additionnelles.

Enfin, chaque sujet du corpus est décrit en plus par une autre variable qualitative que celles qu'on a mentionnées: le parler ou registre utilisé, soit le parler formel, familier ou populaire. On reprend ici les catégories traditionnelles d'un linguiste. Elles sont bien sûr très liées aux autres variables qualitatives disponibles, et il y aurait une étude fine à en faire. Pour simplifier l'exposé, nous ne conservons que le niveau de langage comme variable à expliquer. Notre objectif est donc d'abord de voir si nos variables peuvent séparer les trois types de parlers. Nous voulons également voir comment des variables aussi directement reliées aux propriétés de sélection et d'enchaînement nous permettront de les traduire visuellement. Des analyses factorielles, à la fois sur T_6, et sur les sept premières variables, nous donneront des indications géométriques sur le comportement des locuteurs, et sur l'organisation phonologique du français.

Bien sûr, nous avons utilisé divers tests de validation pour établir que les propriétés mises en évidence ne sont pas des artéfacts originant soit chez les sujets, soit dans la typologie des phonèmes qui a été utilisée, soit dans la classification des sujets selon les trois types de parler. Ainsi pour valider nos résultats avec les types de parlers nous avons généré des types au hasard, et déroulé la même méthodologie. Aucune *ratio* n'est apparue dans les résultats. Les discriminations deviennent impossibles ou très médiocres. Nous avons également engendré des matrices T_6 avec six classes de phonèmes déterminées de façon complètement aléatoire. Mêmes résultats que pour la validation des parlers utilisés. De plus, pour valider finement la typologie des phonèmes, nous avons utilisé divers modèles phonologiques, apportant des variantes dans le contenu de nos six classes de phonèmes. Nos résultats restaient cette fois cohérents, donnant des interprétations factorielles voisines, et nous montrons les résultats avec les meilleures discriminations. Au fond, on pourrait dire que nous apportons ainsi quelque réconfort quantitatif à la typologie retenue... Enfin, de nombreux essais de validation croisée ont montré la stabilité de nos conclusions.

Les résultats des analyses discriminantes que l'on présente plus bas font jouer d'une part les trois groupes, et d'autre part seulement les deux parlers "extrêmes", le formel et le populaire. Le parler familier est rapidement apparu comme une sorte de classe résiduelle, partageant les propriétés des deux classes extrêmes. Les sujets employant ce type de parler se partagent à peu près également dans les deux extrêmes lorsqu'on les place en éléments supplémentaires (comme au dernier tableau de cette section). Les résultats sont à peu près les mêmes que ce soit les variables ou les facteurs qui sont utilisés. Et peu de variables sont nécessaires... Il y a en plus des interprétations possibles à ces discriminations, et ce, en termes linguistiques. Rappelons que nous avons beaucoup utilisé la technique de la validation croisée dans ces analyses, sans pour cela baisser radicalement les taux de bonnes classifications. La suite de nos analyses montrera les performances de nos variables dans la discrimination des parlers extrêmes seulement, les familiers étant placés en éléments supplémentaires.

	Fo.	Fa.	Po.	%
Fo.	28	10	14	54
Fa.	72	31	52	20
Po.	45	24	112	62

	Fo.	Po.	%
Fo.	40	14	74
Po.	122	59	67

Tableau 1. *Matrice de confusion de l'analyse discriminante sur les facteurs. Fo.: groupe formel, Fa.: familier, Po.: populaire. % de bonnes classifications. Les cas de deux et trois groupes.*

Le tableau et la figure qui suivent montrent quantitativement et géométriquement des espaces factoriels, dont les quatre premiers axes expliquent 93 pourcent de la variation totale, et qui sont définis par les variables entropiques ainsi que celles issues de la matrice T_2. On rejoint ici des concepts utilisés depuis longtemps en linguistique, mais on leur donne un sens complètement quantitatif. Il s'agit de ce que SAUSSURE, et après lui beaucoup d'autres linguistes dont JAKOBSON qui l'a utilisé pour établir sa typologie des troubles aphasiques, de la notion de "plan" phonologique où deux axes agissent l'un sur l'autre, l'axe paradigmatique et l'axe syntagmatique.

%	33	27	18	15
variable	comp.1	comp.2	comp.3	comp.4
R_1	-149	-1	886†	242
R_2	70	-101	102	923†
R_3	-91	227	-692†	523
CV	859†	-501	-11	54
VC	-671†	-733†	108	14
λ	-110	986†	81	-52
θ	982†	165	-78	25

Tableau 2. *Composantes principales avec T_2 et les variables entropiques. Pourcentages de variation expliqués par les composantes. Coordonnées factorielles ($\times 10^3$). † montre les principales variables de définition des composantes.*

Nous proposons à vrai dire un plan pour chacun de ces deux aspects naturels du langage. L'axe paradigmatique devient le plan issu des variables de choix, comme il se doit, et l'autre est issu des variables d'enchaînement. On conserve la cohérence. Ces deux plans sont orthogonaux dans leur description factorielle.

Les résultats les plus intéressants du point de vue linguistique concernent les plans factoriels obtenus de la matrice T_6. Nous avons partagé cette matrice en quatre parties décrivant les enchaînements de consonnes à consonnes, de consonnes à voyelles, etc. On les appelle les quadrants CAC, CAV, VAC, VAV. Les analyses factorielles, avec la rotation orthogonale varimax, nous donnent des espaces phonologiques avec des oppositions très marquées dans les divers types d'enchaînements. Le tableau 3 illustre ces oppositions. Par exemple, l'axe 1 associé au quadrant CAV est défini par les oppositions des enchaînements des consonnes de type 1 aux voyelles antérieures et postérieures, et par rien d'autre. Ce remarquable comportement est encore présent lorsqu'on fait l'analyse sur les sujets regroupés par type de parler. Il marque donc une propriété de la phonologie en français montréalais. Les quatre quadrants sont ainsi définis par des oppositions entre types d'enchaînements.

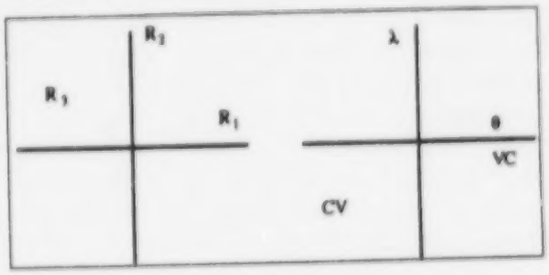


Figure 1. *Projections de variables sur les deux dernières et les deux premières composantes respectivement.*

%	20	20	20	20
variable	comp.1	comp.2	comp.3	comp.4
c_1v_1	956†	-10	10	-49
c_1v_2	-32	80	-37	25
c_1v_3	-954†	-35	11	35
c_2v_1	64	-83	-870†	-454
c_2v_2	-132	117	67	875†
c_2v_3	47	-13	931†	-278
c_3v_1	-26	941†	68	-302
c_3v_2	95	-320	-106	653†
c_3v_3	-57	-844†	17	-243

⇓

SCHÉMA D'OPPOSITIONS

I		-I
-III	IV	III
II	IV	-II

Tableau 3. *Coordonnées factorielles des variables* (× 10^3). *Pourcentages de variation associés aux composantes. Schéma d'oppositions du quadrant CAV.*

Les analyses discriminantes, que ce soit sur les variables elles-mêmes ou sur les coordonnées factorielles donnent sensiblement les mêmes résultats. Seuls les deux quadrants CAC et CAV procurent en fait des discriminations. Et elles sont très stables, compte tenu des techniques de validation utilisées. La signification linguistique de cette séparation des comportements reste en partie mystérieuse. Certains éléments sont décrits dans les références. Il y a sans doute des déterminants lexicologiques dans ces différences, mais que la phonologie, telle que perçue par des variables globales, soit modifiée suffisamment pour séparer les deux types de parlers, c'est un résultat nouveau et intéressant pour les linguistes. La description proposée ici, avec ses schémas d'oppositions, semble prometteuse.

3 Phonétique en sociolinguistique

Traditionnellement la linguistique étudie un locuteur idéal dans une communauté linguistique homogène, et c'est sous l'impulsion de chercheurs américains qu'elle s'est enfin penchée sur la variation de l'usage (LABOV, 1966).

	Fo.	Po.	%
Fo.	38	14	73
Fa.	64	91	—
Po.	51	130	72

Comp.	Fo.	Fa.	Po.
II	706	-10	-171
III	-372	-94	182
I	413	10	-116

variable	Fo.	Fa.	Po.
c_2v_3	251	273	289
c_1v_3	354	388	408
c_3v_2	173	152	149

Tableau 4. *Quadrant CAV. Matrice de confusion. % de bonnes classifications. Valeurs moyennes* (× 10^3) *des facteurs et variables discriminantes sur les groupes.*

Le langage n'est pas qu'un moyen de communication, il sert aussi de signe de reconnaissance sociale, témoigne d'une intégration à un groupe... Le Québec, et Montréal en particulier, est un objet fascinant d'observation pour plus d'une discipline des sciences sociales chez lesquelles, à vrai dire, le langage est toujours prépondérant. Les sociolinguistes ont mis en évidence que le monolinguisme absolu n'existe pas. Chacun de nous utilisons couramment plusieurs dialectes (on dit parfois parlers locaux ou encore registres), parce que nous appartenons, suivant les moments de la journée ou de sa vie, à plusieurs groupes selon qu'on est au travail, dans la famille, selon qu'on est un jeune enfant, un adolescent en révolte, etc. Chaque intégration provoque des variantes du langage maternel (ou d'un autre ...). L'évolution de ces parlers fait aussi l'objet d'investigation, cela va de soi, et, évidemment, l'Analyse des données est constamment appelée à servir.

En 84, un sous-ensemble des sujets du corpus SC fut de nouveau échantillonné, et depuis peu on commence à voir des études sur l'évolution du français à Montréal (YAEGER, 1987 et 1988). Madame Yaeger, a mesuré, —c'est un travail de bénédictine!—, les principaux paramètres des fréquences en Hertz de certaines voyelles dans leur enchaînements lorsqu'énoncées dans des contextes spécifiques. Ces mesures furent prises dans les deux corpus SC, et la question se pose de quantifier les différences de 71 à 84. Des hypothèses sociolinguistiques sont alors mises à l'épreuve.

Le problème toujours présent dans ce type de données est leur petit nombre. De plus, il faut trouver des mesures synthétiques qui dérivent des paramètres mesurés effectivement, et qui permettent des comparaisons. Dans ce qui suit, nous décrivons brièvement les variables et la méthodologie utilisée. Il va de soi que ces études en sont encore à leur tout début, leur exploitation nécessitera plusieurs années. L'analyse factorielle sera, ici aussi, mise à contribution. Nous parlerons de quelques différences mises en évidence à ce jour, et esquisserons un certain nombre de travaux potentiels en se basant sur quelques tableaux factoriels illustratifs.

Le phonéticien utilise les sonogrammes pour décomposer les sons du langage selon l'intensité des fréquences présentes dans son spectre (figure 2, tirée de LADEFOGED, pp. 170-171). Une longue procédure, aujourd'hui en bonne partie informatisée, permet de recueillir pour chaque voyelle observée les principaux paramètres de cette décomposition. Entre autres paramètres, on a conservé les valeurs des deux premières fréquences présentes au-dessus de la fondamentale, et ce au début et à la fin des énonciations de certaines voyelles dans certains contextes: $F_{1,d}$, $F_{2,d}$, $F_{1,f}$, $F_{2,f}$. La fondamentale est due aux cordes vocales, on la néglige, car seuls nous intéressent les deux premiers *formants*, qui colorent les sons par les fréquences de résonance du conduit vocal. Nous ne parlerons que de la voyelle [ɛ] (comme dans 'lait') dans les contextes de famille ou non pour l'enchaînement à un [r]: frère, mère, salaire, hiver, etc. Dans ces contextes on retrouve beaucoup de diphtongaisons. On se retrouve ainsi avec quatre catégories libellées: ERFD ET ERF, ERD et ER, où le F dénote la famille et D la diphtongaison.

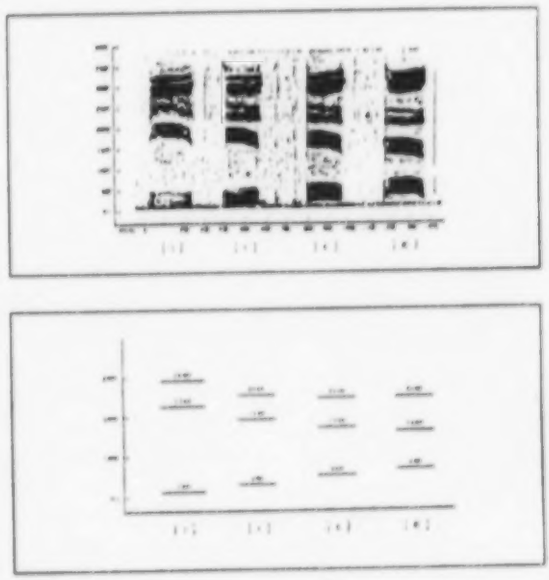


Figure 2. *Sonogrammes des mots anglais: 'heed','hid, 'head', 'had'; fréquences des trois premiers formants. Prononciation en anglais RP (pour received pronounciation, le bon parler anglais...), celui des universitaires, de la BBC, etc.)*

À un niveau plus général, on peut considérer le même phonème suivi par d'autres consonnes, ou dans des contextes plus spécialisés. Il y a bien sûr les autres voyelles et les liens entre les quatre variables conservées selon les types, les contextes. Tout cela s'inscrit ici dans la durée, et on s'intéresse avant tout à la variation au cours du temps de la phonétique des locuteurs.

Le tableau qui suit montre l'évolution de plusieurs des enchaînements que l'on a conservés. Les traits marquent le passage de 71 à 84; seules les évolutions aux traits pleins sont statistiquement significatives. Nous utilisons une applica-

tion simple de l'analyse factorielle pour quantifier le terme de significatif de la façon suivante.

On se restreint d'abord aux seules fréquences initiales. Une analyse factorielle sur ces deux variables en utilisant les mesures seulement pour l'enchaînement considéré, et ce dans les deux années échantillonnées nous donne une variable synthétique, et un test-t usuel sur les coordonnées factorielles décide de la significativité de l'écart entre les deux années. Le niveau du cinquième percentile décide du trait plein.

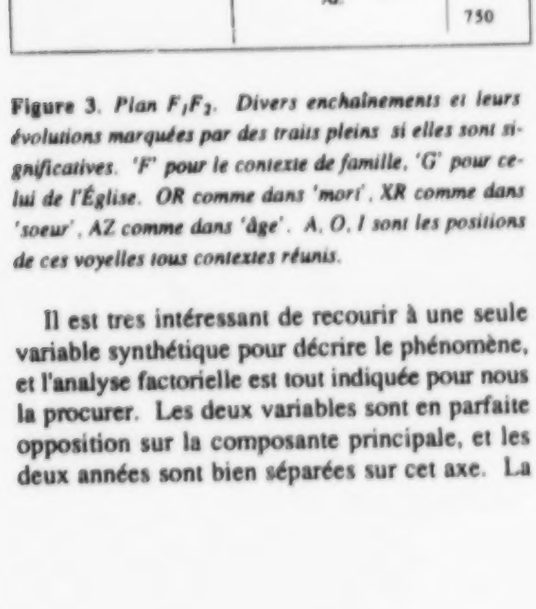
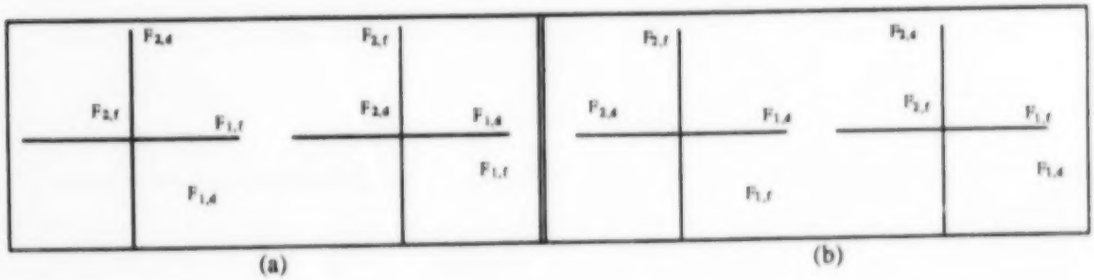


Figure 3. *Plan F_1F_2. Divers enchaînements et leurs évolutions marquées par des traits pleins si elles sont significatives. 'F' pour le contexte de famille, 'G' pour celui de l'Église. OR comme dans 'mort', XR comme dans 'soeur', AZ comme dans 'âge'. A, O, I sont les positions de ces voyelles tous contextes réunis.*

Il est tres intéressant de recourir à une seule variable synthétique pour décrire le phénomène, et l'analyse factorielle est tout indiquée pour nous la procurer. Les deux variables sont en parfaite opposition sur la composante principale, et les deux années sont bien séparées sur cet axe. La composante principale décrit la *couleur* de la voyelle énoncée. Toutefois, nous avons de sérieux problèmes de validation. Ainsi, nous sommes souvent à la frontière du viol des conditions d'application du test-t. Le relatif petit nombre d'observations donne des problèmes de stabilités aux moyennes, écarts types, corrélations, et aux ... statistiques.

On perçoit clairement cependant des différences phonétiques année sur année. Des variations opposées de certains enchaînements montre que le langage a changé, pas seulement la voix, et, à ce jour, on observe des tableaux analogues au précédent pour six sujets. Se pose maintenant le problème de la normalisation des sujets, pour ramener les sujets au même diapason en quelque sorte; autrement pas de comparaison possible. La fonction, s'il y en a une (peut-il en être autrement?), de cette variation du langage est en grande partie mystérieuse. Mais mouvement il y a ... *eppur si muove*, ce qui, il y a à peine quelques années, faisait l'objet d'hypothèses à l'opposé, et était purement spéculatif. Nous renvoyons le lecteur sérieux aux références pour le détail des considérations linguistiques qu'il est impensable de rapporter ici! Le phénomène, chose certaine, est cerné de plus en plus près.

Nous terminons en montrant les principaux plans factoriels calculés sur les données d'un seul sujet et de tous les sujets pour un type d'enchaînements, celui de ERF. L'exploitation des plans factoriels d'enchaînements est à peine entamée. On ne sait jamais, cela peut aboutir quelque part...



(a) (b)

Tableau 5. *Plans factoriels 1-2 et 1-3 de ER. (a) Un seul sujet et (b) tous les sujets réunis à ce jour.*

4 Conclusion

Depuis les quinze dernières années, et ce grâce à la disponibilité accrue de puissants moyens de traitement de l'information, on a vu un développement remarquable de la linguistique quantitative. On est passé de l'étude, somme toute assez primitive, de modèles d'abondance, presque la seule préoccupation quantitative en linguistique, à une floraison de "modèles" explicatifs de quelque phénomène. Nous écrivons modèles entre guillemets parce qu'il est encore trop souvent difficile de voir la pertinence linguistique de beaucoup d'études. Les modèles retrouvent les caractéristiques sociologiques des sujets — "à la Higgins", le linguiste du Pygmalion de G. B. Shaw! — mais qu'est-ce qui est vraiment en jeu dans la langue des locuteurs? C'est en fait la question la plus importante. Plus généralement, que peut apporter de nouveau l'analyste des données aux diverses disciplines qu'il sert?

En ce qui concerne la linguistique, il apparaît clair qu'il faudra dépasser la cadre des tableaux de contingence et de leur analyse, selon le paradigme méthodologique de Benzécri que nous avons mentionné au début. Il importe, avant tout croyons-nous, de trouver des instruments de mesure directement descriptifs de propriétés linguistiques, et que les interprétations, ou modèles produits, aient un contenu linguistique prégnant. Ceci semble le cas plus nettement dans le modèle phonologique que nous avons exposé. L'autre exemple d'analyse factorielle dont nous avons parlé aussi montre en fait encore plutôt une astuce de calcul qu'un développement de modèle à proprement parler.

Le modèle phonologique que nous avons exposé en premier pourrait éventuellement permettre d'observer des évolutions dans les structures phonologiques des individus, que ce soit dans les cas de perturbation, de récupération, ou même d'apprentissage. On peut y apporter des extensions, à la syntaxe par exemple, ou encore se demander quelles sont les caractéristiques stylistiques qu'il permettrait de mesurer. On peut évidemment songer à des applications aux diagnostics des aphasies. Ceci a d'ailleurs fait l'objet du premier article que nous avons écrit sur le sujet. Il nous manquait alors un bon ensemble d'apprentissage. Ce problème est aujourd'hui en bonne partie réglé. Il reste à voir quels développements seront apportés.

Quoiqu'il en soit, il est bien clair que nous sommes entrés dans une ère nouvelle, où une espèce de prolongement du cerveau dans des réseaux d'information ou de procédures creusés dans le silicium, vient apporter des possibilités radicalement nouvelles, encore à peine soupçonnées. L'ordinateur a une nature en fait intrinsè-

quement linguistique, comme son géniteur, d'ailleurs, avec son intelligence dite naturelle...

Références

†AIT-HAMLAT A., 1984: Analyse des répétitions et indexation automatique des documents. CAD, vol. IX, no. 2, pp. 173-204.

AIT-HAMLAT A.., 1983: Analyses comparatives d'un ensemble de données lexicales. CAD, vol. VIII, no. 2, pp. 199-204.

AZAR M., 1970: Analyse morphologique du texte hébreux de la bible, 2 tomes. Cahiers du CRAL, no. 12 et 13, Nancy.

BEAUDRY M., 1978: Étude du signal vocal dans sa représentation amplitude-temps; algorithme de segmentation et de reconnaissance de la parole. Thèse de 3e cycle. Université de Paris VI.

BEHRAKIS T., 1983: Vers une analyse automatique des textes; le traitement de 42 observations contenues dans les livres Épidémies I et III d'Hyppocrate. CAD, vol. VIII, no. 3, pp. 475-489.

BENZÉCRI J.-P., 1984: Description des textes et analyse documentaire. CAD, vol. IX, no. 2, pp. 205-211.

BENZÉCRI J.-P., 1986: Élaboration statistique de données sur ordinateur; application à l'analyse des textes; contributions attendues à l'étude de la bible. CAD, vol. XI, no. 1, pp. 7-30.

BENZÉCRI J.-P., FÉGHALI C.H., 1983: L'articulation des voyelles; analyse sur micro-ordinateur et interprétation des résultats avec ceux issus d'une analyse ternaire. CAD, vol. VIII, no. 2, pp. 159-180.

BENZÉCRI J.-P., éditeur, 1981: Pratique de l'analyse des données (tome 3); linguistique et lexicologie. Paris: Dunod.

BERGHAUSS F.G.: Remarques sur la validité de différentes méthodes de groupement automatique visant à éclaircir les rapports entre les manuscrits. *In* IRIGOIN, *op. cit.*, pp. 97-111.

BOURDEAU M., 1986: On some quantifications in phonology. Applications to deviant language productions. ITL Journal of applied Linguistics, 74, pp. 17-45.

BOURDEAU M., DAGENAIS L., SANTERRE L., 1988: Quantitative frames for the phonological plane. Soumis pour publication.

BRAMUCCI F., COSTA P., MARAVELLE M., 1982: Actif et moyen dans le verbe. Analyse de données nouvelles recueillies en Italie en réponse à un questionnaire. CAD, vol. VII, no. 3, pp. 372-377.

CARRETERO J., 1979: Mesures de propriétés statistiques de textes en langue naturelle pour la définition de modèles linguistiques utilisables en analyse morpho-syntaxique. Thèse de l'Université de Lyon I.

CHARPILLET F., 1985: Un système de reconnaissance de la parole contenue pour la saisie de textes lus. Thèse de doctorat de 3e cycle, Université de Nancy I.

CHERRY E.C., HALLE M., JAKOBSON R., 1953: Toward the logical description of languages in their phonemic aspect. Language 29, pp. 449-463.

DUPLACY J., 1977: Classification mathématique et informatique des états d'un texte: repères historiques et recherches méthodologiques. Revue d'histoire des textes, Éditions du CNRS, tome 5 (1975), pp. 249-309, Paris.

FARAH A. A., 1986: Un estimateur du lexique d'un texte de situation. RSA, XXXIV, no. 2, pp. 63-75.

GALLOWAY P., 1979: Manuscript filiation and cluster analysis, the "Loi de l'ombre" case. *In* IRIGOIN, *op. cit.*, pp. 87-95.

GRIFFITH J. G.: Non systematic classification of manuscripts by computer methods. *In* IRIGOIN, *op. cit.*, pp. 79-86.

HATTON J.-P., LAMOTTE M., 1971: Étude statistique des phonèmes et diphonèmes dans

† Abréviations: CAD, Cahiers de l'Analyse de données; RSA, Revue de Statistique Appliquée.

le français parlé. Revue d'acoustique, no. 16, pp. 253-262.

IRIGOIN J., ZARRI G. P., 1979: La pratique des ordinateurs dans la critique des textes. Paris: Éditions du CNRS.

JAKOBSON R., 1964: Towards a linguistic typology of aphasic impairments. *In* Disorders of language, A.V.S. de Reuck et M. O'Connor éds. London: Churchill, pp. 21-42.

LABOV W., 1966: Stratification of English in New York City Speech. Washington D.C.: Center for Applied Linguistics.

LADEFOGED P., 1975: A course in phonetics. New York: Harcourt Brace Jovanovich.

LEBART L., 1981: Une procédure d'analyse lexicale écrite en fortran. CAD, vol. VI, no. 2, pp. 229-243.

LEBART L., 1969: Analyse statistique de la contiguïté. Publication de l'ISUP, vol. XVIII, Paris, pp. 81-112.

LELU A., ROSENBLATT D., 1986: Représentation et parcours d'un espace documentaire. Analyse des données, réseaux neuronaux, et banques d'images. CAD, vol. XI, no. 4, pp. 453-470.

LERMAN I. C., 1981: Classification et analyse ordinale des données. Paris: Dunod.

LERMAN I. C., PETER P. H., 1985: Organisation et consultation d'une banque de petites annonces à partir d'une méthode de classification hiérarchique en parallèle. Actes des 4e Journées Int. d'analyse des données et informatique, Inria, Rocquencourt.

MANDELBROT B., 1961: On the theory of word frequencies and on related markovian models of discourse. *In* R. JAKOBSON éd., Proc. Symp. Appl. Math., vol 12., pp. 190-219.

MARKOV A. A., 1913, réimpression en 1964: Un exemple de recherche statistique sur le texte d'Evgenii Onegin illustrant la liaison des épreuves en chaînes. Bulletin de l'académie royale des Sciences, 6e série, t.7, pp. 153-162.

MORIN A., 1981: Recherche d'une métrique optimale en analyse de données. Application en lexicométrie. Thèse de doctorat de 3e cycle, Université de Rennes I.

MORIN A., 1984: Comparaison de plusieurs méthodes de classification sur un exemple de lexicométrie. RSA, vol. XXXII, no. 4, pp. 37-49.

MZABI M., 1983: Le mot "joie" dans 8 romans en langue d'oïl du XIIIe. CAD, vol. VIII, no. 3, pp. 332-342.

PÉCHEUX M., 1969: Analyse automatique du discours. Paris: Dunod.

REINERT A., 1983: Une méthode de classification descendante hiérarchique; application à l'analyse lexicale par contexte. CAD, vol. VIII, no. 2, pp. 187-198.

REINERT M., 1986: Un logiciel d'analyse lexicale, CAD, vol. XI, no. 4, pp. 471-484.

ROUSSEAU P., SANKOFF D., 1979: Analyse typologique de données binomiales; application en linguistique. E. Diday éditeur, Optimisation en classification automatique, tome I, INRIA, pp. 429-442.

SALEM A., 1984: La typologie des segments répétés dans un corpus, fondé sur l'analyse d'un tableau croisant mots et textes. CAD, vol. IX, no. 4, pp. 489-500.

SANKOFF D., 1987: Variable rules. Rapp. Tech. CRM-1450, Centre de recherche en mathématiques appliquées, Université de Montréal.

SANKOFF D., LABERGE S., 1978: The linguistic market and the statistical explanation of variability. *In* D. Sankoff éd., Linguistic variation: Models and methods. New York: Acad. Press, pp. 239-250.

SANKOFF D., LESSARD R., TRUONG N.B., 1978: Computational Linguistics and Statistics in the analysis of the Montreal French corpus. Computers and the Humanities 11, pp. 185-191.

SANKOFF D., SANKOFF G., LABERGE S., TOPHAM M., 1976: Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale. Cahier de Linguistique no. 6, Montréal: Presses de l'Université du Québec, pp. 85-125.

SHANNON C.E., 1948: A mathematical theory of communications. Bell System Tech. Jour., vol. 27, pp. 379-423 et 623-656.

TUBACH J.P., 1970: Reconnaissance automatique de la parole, étude et réalisation fondées sur les niveaux acoustiques; morphologie et syntaxe. Thèse de doctorat de 3e cycle, Université de Grenoble I.

WEIL G.E., 1974: Bible hébraïque et Targum araméen. Traitement automatique et étude quantitative. *In* P. Flajolet éd., informatique et philologie, colloque INRIA, pp. 5-39.

YAEGER-DROR M., BOURDEAU M., 1978: Montréal French vowel duration. Jour. Acoustical Soc. Amer., vol. 63, S87.

YAEGER-DROR M., 1987: Changement en chaîne dans le français parlé à Montréal. *In* D. Sankoff éd., Diversity and Diachrony, pp. 222-238. Philadelphia: Benjamins.

YAEGER-DROR M., 1988: Real time change in Montréal French, Sociolinguistics Symp. 7. To be published in the York Papers in Linguistics, University of York, UK.

YAGLOM A.M., YAGLOM I.M., 1969: Probabilité et information. Paris: Dunod.

MODÉLISATION STATISTIQUE D'UNE APPROCHE SYSTÉMIQUE: CRISE SCOLAIRE, MÉTAPHORES VERBALES ET ANALYSE DES CORRESPONDANCES

Michelle Bourassa
Faculté d'éducation, Université d'Ottawa

1 Introduction: Pour une vision systémique d'une crise

"La vraie objectivité ne réside pas dans l'absence du sujet mais dans la nature qu'il entretient avec son objet. Il est grand temps de désobjectiver la science pour la libérer de l'obsession de la subjectivité, non pas pour la rendre subjective... C'est donc la révolution keplérienne que l'on évoque ici. Son originalité, on s'en souvient, a mis le soleil, devenu point moteur effectif, en un des foyers de la courbe que les planètes décrivent autour de lui, courbe variable et non plus fixe, qui dépend des forces s'exerçant entre le corps central et les planètes qui en dépendent. En épistémologie, cela signifie que le sujet, son observation, ses catégories de pensée, occupant un des foyers du système de connaissance détermine le ciel observé, les phénomènes qui s'y déroulent et qui, pas plus que les corps célestes, ne parcourent une orbite fixe ni sont sans rapport les uns aux autres." Edgar Morin.

À titre de psychologue, je me retrouve régulièrement confrontée à une situation chronique et irréversible, celle de la projection. Projection de la faute [qui assume la *cause originelle* d'une difficulté]; projection du traitement [à qui incombe ultimement la responsabilité du traitement]. Toute projection sous-tend un double refus: refus de reconnaître tout point de vue autre et refus de reconnaître qu'il y a refus. La nécessité d'affirmer sa vision comme unique et intransitive transforme tout argument en discours positiviste; le modèle devient causal non-réciproque — faire la preuve — plutôt que développemental — expliquer pour apprendre.

Cette recherche étudie l'évolution des relations de projection entre quatre familles et un service scolaire pour enfants présentant des troubles graves de communication. En tant que microcosme, d'image réduite de la réalité, elle permet d'identifier la dynamique évolutive de systèmes humains *condamnés* à interagir dans une relation verticale aidé/aidant. C'est en traitant directement les données verbales tirées des entrevues consignées durant les quatre années de ce service qu'il devint possible de saisir la structure épigénétique d'une relation d'aide et ses limites...

Tout traitement de données impose ses propres contraintes méthodologiques. Lorsque ces données proviennent d'entrevues, de telles contraintes se trouvent automatiquement décuplées par leur nature comme leur quantité. Si l'on ajoute à cela la conviction épistémologique que la validité des résultats est directement tributaire de la validité de construit des entrevues, il s'ensuit d'une part que celles-ci ne peuvent qu'être ouvertes et d'autre part, qu'il sera conséquemment impossible pour le chercheur d'imposer une classification a priori des données [d'imposer un canevas fixe d'entrevues].

Cette situation suppose une méthode d'analyse qui puisse simultanément traiter des informations non-ordonnées et permettre une mise en relation individus/caractéristiques. L'analyse factorielle de correspondance répondait à cette double-contrainte. En découvrant les logiciels AFC* ACP et UTIL (version 3.0) créés pour les systèmes MACINTOSH par JEAN BRUNET [départe-

École d'été sur l'analyse des données, juin 1988.
SCAD-CIRADE

ment de communication, U.Q.A.M.], j'entrevis la possibilité d'effectuer une analyse à la fois puissante et rigoureuse. Il me restait à identifier, dans le corpus de mes données, des caractéristiques qui se prêteraient à ce type d'analyse.

De multiples relectures des entrevues m'amenèrent à prendre conscience des innombrables redondances verbales qui truffaient l'ensemble des corpus. Ces redondances ou *métaphores* verbales touchaient les définitions que mon milieu de recherche donnait des enfants desservis par le service scolaire, du service scolaire lui-même et finalement des comportements interactifs entre les parents et le service scolaire. Trois listes de données ont conséquemment été consignées: 1) la liste des métaphores sur l'enfant, 2) celle sur l'école et 3) celle définissant les comportements des deux partis. Ces listes permettaient de distinguer quels partenaires jugeaient quelles métaphores acceptables ou inacceptables. Elles furent donc traduites en tableaux booléens distinguant deux modalités: la modalité décrivant le présent et celle décrivant l'état attendu. Lorsque la même caractéristique était utilisée pour décrire les deux modalités, on pouvait conclure que les partenaires de cette recherche la jugeaient acceptable; lorsque le cas contraire se présentait, on devait conclure qu'un changement était attendu.

Les avantages de l'analyse factorielle des correspondances tiennent non seulement dans le fait qu'elle est capable de traiter le type particulier de données de cette recherche mais aussi dans le fait qu'elle fournit une carte perceptuelle permettant d'appréhender d'un seul coup d'oeil l'ensemble des informations. Il devenait donc possible d'envisager présenter les résultats de cette recherche aux partenaires eux-mêmes.

L'analyse de ces associations sémantiques dessinées sur cartes perceptuelles permit deux découvertes: d'abord, que deux systèmes apparemment en contradiction maintiennent la chicane non pas parce qu'aucun recoupement dans leurs modes de pensée n'est possible mais au contraire parce qu'un méta-mode illicite, donc innommable, les réunit. Ce méta-mode parle du concept d'*utopie* — soit qu'une solution définitive [en l'occurence guérir l'autisme des enfants desservis] est simultanément nécessaire [il faut guérir ces enfants si le système veut survivre] et impossible [l'autisme n'est pas guérissable]. L'étude révèle de plus que l'utopie se maintient grâce au fait que l'attention des acteurs se trouve déviée vers des problèmes secondaires. Le principal problème secondaire, c'est l'expression de dissidences concernant la distribution du pouvoir. Dans tout système expert, il est inévitable qu'ayant posé la prémisse d'expertise ["il existe une solution finale"], on attende de l'autre la résolution. Un tel modèle permet non seulement d'instaurer la chicane face à la distribution du pouvoir et la projection comme mécanisme de défense face à toute controverse mais aussi de continuer à ignorer collectivement l'impossible — rééduquer l'enfant autistique.

Les clients de toute aide doivent, pour l'accepter, admettre au départ leur incompétence à résoudre eux-mêmes leurs difficultés. Cette aide implique une confiance absolue dans l'expertise de l'autre de sorte qu'à moyen terme, elle entraîne l'émancipation de l'aidé. Or, l'utopie de la cure de l'autisme, révélée par l'analyse des correspondances des métaphores sur l'enfant, rend caduque tout espoir d'une émancipation définitive. Trois groupes d'intervenants se trouvent dès lors également piégés: les enfants, les parents, le service scolaire. C'est l'évolution d'un tel piège se refermant inéluctablement sur l'ensemble des participants qui fait l'objet de cette conférence.

2 Problématique

2.1 Caractéristiques fondamentales des systèmes souples

Tout système souple agit comme un organisme en développement, se renouvelant de lui-même sans jamais atteindre l'état final ou ultime. Il reste informé de ses progrès vers ses objectifs de croissance grâce à un processus d'anticipa-

tion-rétroaction. Advenant que ce processus soit positif, le système, mis en confiance, utilisera ses structures dissipatives pour accroître ses échanges avec l'environnement et ainsi sera susceptible d'atteindre un nouvel état — *autopoesis* — à équilibre plus complexe (PRIGOGINE, 1972, BRENT, 1978). S'il est négatif, le système menacé se recalibre par prolifération de structures de contrôle; ces structures entraînent une régression vers un état d'équilibre antérieur, dit homéostatique (BATESON, 1978).

Lorsqu'il y a une quantité excédentaire de structures homéostatiques par rapport aux structures dissipatives, le système perd la *variété requise* pour assurer son développement (ASHBY dans ROWEN & REASON, 1982). La dynamique protectionniste qui en résulte prévient toute rééquilibration vers un état supérieur. Pour échapper à cette rigidification, le système doit permettre la mise en place de nouveaux mécanismes de communication avec l'environnement. Ces mécanismes assurent le développement de nouvelles structures participant à un niveau d'organisation supérieur. "Il suffirait de considérer l'organisation comme un processus ininterrompu de désorganisation-organisation, et non pas comme un état..." (ATLAN, 1979: 45) pour que la recherche de l'équilibre ne soit plus qu'une utopie nécessaire à la croissance des systèmes souples et que l'enjeu équilibre-déséquilibre devienne partie intégrante de tout système souple.

2.2 Caractéristiques fondamentales du système étudié

Cette recherche traite des tentatives de croissance de deux systèmes humains intimement reliés, un service scolaire pour enfants autistiques et les quatre familles de ces enfants. Alors que le service scolaire présente une position unanime, les familles se perçoivent comme formant trois catégories distinctes. Deux d'entre elles, parce que difficilement dissociables du milieu scolaire tant par leurs propos que par leurs actions, sont qualifiées par l'auteur de familles symbiotiques. Une troisième famille, contestant l'approche du service mais incapable de se résoudre à retirer son enfant, est qualifiée d'ambivalente. Enfin, une dernière, fort vindicative et éventuellement dissociée du service, est qualifiée de dissidente.

Une telle classification laisse supposer que la vision de la situation problématique comme des solutions à apporter diverge considérablement d'une famille à l'autre. Tous les partenaires en sont en fait convaincus, chacun tentant à sa façon de *convertir* ces autres qui ne veulent rien comprendre. Ces tentatives unidirectionnelles ont comme conséquences inévitables que ceux qui doivent changer ne sont jamais ceux qui définissent le problème ni sa solution. Projection donc, de la cause (du problème) comme du traitement (de la solution).

Fidèle au modèle traditionnel expert-aidé, cette projection est historiquement initiée par le service scolaire. La difficulté ne provient pas du fait que ce service propose sa vision mais du fait qu'il l'impose comme seule valide. Trois conséquences en découlent: on se retrouve avec un problème de validité de construit *(définition et solution tombent du ciel);* la distribution des rôles est déterminée *de facto* et considérée invariante; l'enjeu projectif devra éventuellement être assumé par les aidés dans une tentative souvent désespérée d'émancipation d'une histoire sans issue.

Confondant toutes les théories en résolution de problème stipulant que rien ne va plus lorsque les partenaires n'arrivent pas *à regarder ensemble dans la même direction*, l'analyse des correspondances, en mettant en évidence les innombrables associations sémantiques *communes* à l'ensemble des acteurs, impose la thèse inverse. Les difficultés relationnelles ne proviennent pas de visions irréconciliables, impossibles à consensualiser mais de visions si totalement identiques et redondantes que le circuit en est aveuglé. L'absence de vision contrastante devient une nouvelle forme de perte de *variété requise*.

"Binocular vision works not because the two eyes see two different things but because the differential of the two images enables the brain to compute the invisible dimension," (MARUYAMA, 1983: 598). L'analyse des correspondances m'a tenu lieu de vision binoculaire. Ses caractéristiques particulières m'ont appris à reconnaître les symétries qui recouvrent les divergences apparentes dans la vaste majorité des interactions humaines.

3 Analyse des correspondances

3.1 Choix épistémologiques

Cinq facteurs justifient le choix de ce type d'analyse dans le traitement des données. Il y a d'abord leur immense complexité occasionnée par la quantité (quatre années de recherche participative) et le type d'informations recueillies (verbatim d'entrevues ouvertes et semi-ouvertes) ainsi que leurs innombrables recoupements (redondances des informations). Une voie de solution m'est apparue alors que je me laissais habiter par la relecture de ces entrevues. Les métaphores verbales — imageries mentales servant de synthèses définitionnelles de la situation — s'imposèrent à mon insu soit parce qu'omniprésentes dans l'ensemble des entrevues ou encore parce que centrales pour un sous-ensemble. Ces métaphores en vinrent donc à constituer une longue suite de caractéristiques qui pouvait désormais être mise en correspondance avec les divers ensembles d'individus (école, familles symbiotiques, ambivalente, dissidente).

Il y a ensuite l'épistémologie propre à l'analyse multidimensionnelle. Ce type d'analyse suppose arbitrairement que les ensembles d'individus et de caractéristiques qui leur sont attachées n'ont aucune structure a priori, constituées uniquement de suite d'éléments. Ce traitement en concaténation respecte un postulat fondamental de cette recherche à savoir que le rapport entre éléments n'est jamais linéaire et causal, mais circulaire et interactif. De plus, il me permet de réduire les risques d'un biais face aux données puisque l'ordonnancement ne peut être postulé donc justifié qu'a posteriori.

Il y a aussi l'avantage de pouvoir utiliser les tableaux booléens pour formater les données initiales. Ce genre de tableau est non seulement statistiquement très stable mais il permet de comparer simultanément deux modes de traitement. Or, les métaphores verbales répertoriées distinguent explicitement entre le mode *présent* et le mode *anticipé*. Le tableau booléen rendait possible la compilation simultanée du portrait considéré réel (mode présent) et du portrait considéré idéal (mode attendu).

Il y a encore le fait que l'analyse fournit une représentation sous forme de carte perceptuelle permettant d'appréhender d'un seul coup d'oeil l'ensemble des données. Créer une carte perceptuelle permettant de voir d'un seul coup d'oeil le jeu interactif des images verbales... grâce à la mise en relation multi-axiale nous permet de dégager une vision d'ensemble. Considérons deux termes servant à définir les enfants, *anormal* et *déviant*. Le Petit Robert les définit respectivement comme "ce qui n'est pas conforme aux règles et lois établies" et "tout comportement qui s'éloigne de la norme sociale." On pourrait conclure que ces termes sont synonymes et qu'ils occupent conséquemment une position confondue sur les cartes perceptuelles. Il n'en est rien. L'étiquette *anormal* appartient au nuage des caractéristiques intolérables et inacceptables de l'enfant alors que l'étiquette *déviant* appartient au nuage des caractéristiques tolérables donc acceptables. Seule l'analyse des correspondances pouvait distinguer ce qui, intuitivement, aurait été confondu. Il faut conclure que toute métaphore est plus que son image et qu'elle ne prend de sens que mise en relation avec les autres métaphores propres à une population donnée.

Il y a enfin le jeu de va-et-vient continu entre les divers modes d'interprétation, jeu qui place les relations de dualité au centre de la discussion. Les tendances centrales (*représentées par les axes orthogonaux d'inertie, ces tendances décrivent l'ensemble des*

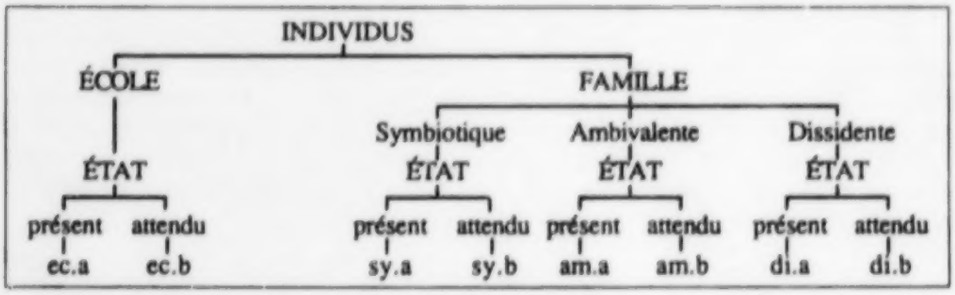
caractéristiques en traits dominants ou facteurs) sont ainsi comparées aux tendances de dispersion (*représentées par l'écart entre les axes, ces tendances établissent la relation entre les facteurs*); les points à corrélation forte proches du centre de gravité sont comparés aux points éloignés même s'ils n'ont qu'une corrélation faible. En procédant de la sorte, on diminue tout risque d'interprétation abusive; l'information tirée de l'un se trouve sans cesse pondérée par l'information tirée de l'autre. De plus, en maintenant dans l'analyse les points périphériques (éloignés du centre de gravité), il devient possible de comprendre des variations qui seraient demeurées inexplicables si l'on n'avait considéré que les tendances centrales. La discussion qui suivra illustrera clairement ces aspects.

3.2 Type de données

Deux types de données seront mis en correspondance:

- l'ensemble des individus réparti en quatre sous-ensembles: école, familles symbiotiques, famille ambivalente, famille dissidente; chacun de ces sous-ensembles est subdivisé en deux sections, la section des données qui décrit l'état présent tel que perçu, représenté par la lettre *a* ainsi que la section des données définissant l'état tel qu'attendu, représenté par la lettre *b*. Le tableau 1 synthétise la répartition de la population étudiée.

Tableau 1: Répartition de la population étudiée



- l'ensemble des caractéristiques constitué des métaphores utilisées par les intervenants pour dépeindre le contexte se scinde en deux types: les caractéristiques définissant les objets d'étude, *l'enfant* et *l'école* et les caractéristiques définissant les enjeux de la crise, soit *les comportements* des intervenants impliqués. Le tableau 2 résume ces caractéristiques.

Tableau 2: Caractéristiques des objets d'étude

CARACTÉRISTIQUES			
MÉTAPHORES DÉFINITIONNELLES		MÉTAPHORES COMPORTEMENTALES	
ENFANT	ÉCOLE	SYMÉTRIQUES	COMPLÉMENTAIRES
(27)	(11)	(9)	(3)

Il est à noter que chacune des caractéristiques est traitée deux fois, d'abord dans le mode actuel, ensuite dans le mode attendu.

Le nombre limité de métaphores retenues s'explique d'une part par leur rôle central et d'autre part, par la possibilité d'opérer une confirmation de la validité des résultats en examinant chaque métaphore dans une variété de mises en relation. L'important dans ces analyses n'est pas de parvenir à une explication intransitive et définitive mais de découvrir l'unité de signification commune à l'ensemble des éléments. De telles *découvertes* servent à donner un sens aux choses. En reliant les métaphores, les analyses suppriment leur caractère surprenant; en distinguant celles qui ne doivent plus être confondues, elles préservent leur caractère unique. Il devient dès lors possible de prédire le type d'associations le plus probable à un moment donné de l'histoire.

3.3 Tableaux et graphiques

- TABLEAUX:

Tous les tableaux booléens qui suivent, expriment la fréquence de relation entre les individus et les caractéristiques, fréquence représentée sous forme disjonctive où l'appartenance égale 1, la non-appartenance, 0. Les colonnes illustrent la distribution des caractéristiques, les rangées, la distribution des individus.

Tableau 3. Termes retenus

1 différent (diff)	8 égocentrique (egoc)	15 incontrôlable (inco)	22 comprend tout (c.t.)
2 anormal (anor)	9 sourire (sour)	16 symbolique (symb)	23 médium (medi)
3 déviant (devi)	10 contact visuel (c.v.)	17 énigmatique (enig)	24 dévorant (devo)
4 imprévisible (impr)	11 joue seul (j.s.)	18 écholalique (echo)	25 orientation académique (o.a.)
5 violent (viol)	12 joue avec fatrie (j.f.)	19 pleurnichard (pleu)	26 intégration académique (i.a.)
6 hystérique (hyst)	13 joue avec pairs (j.p.)	20 paresseux (pare)	27 intégration sociale (i.s.)
7 colérique (cole)	14 joue avec adultes (j.a.)	21 ne veut pas (n.v.)	

Tableau 4. Métaphores sur les enfants

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	26	27
	Termes																									
s.a	1	0	1	1	1	1	0	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	1	0	0	0
s.b	1	0	1	0	1	1	0	1	0	1	0	1	1	0	1	0	0	0	0	0	1	0	0	1	1	1
a.a	1	1	1	1	0	1	1	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0
a.b	1	0	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	1
d.a	1	0	0	0	0	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1
d.b	1	0	0	0	0	0	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	1	1	1
e.a	1	0	1	1	1	0	1	1	1	1	0	0	1	0	1	1	1	1	0	0	0	0	1	0	0	0
e.b	1	0	1	1	0	0	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1

Tableau 5. Métaphores sur l'école

1	expertise (exp)	8	respect des particularités (resp)
2	évaluation (anoi)	9	enfant comme personne (enf)
3	innovation (devi)	10	parent-substitut (par)
4	contrôle (cont)	11	oreille attentive (orei)
5	conseiller (cons)		
6	ouverture d'esprit (ouve)		
7	intégration (inte)		

Tableau 6. Métaphores sur l'école

	1	2	3	4	5	6	7	8	9	10	11
sy.a	1	1	1	1	1	1	0	1	1	1	1
sy.b	1	1	1	1	1	1	1	1	1	1	1
am.a	1	1	0	1	0	0	0	0	1	1	1
am.b	1	1	1	1	1	0	1	1	1	1	1
di.a	0	1	0	1	1	0	0	0	0	0	1
di.b	1	1	1	0	0	1	1	1	1	1	1
ec.a	1	1	1	1	1	1	0	1	1	1	1
ec.b	1	1	1	1	1	1	1	1	1	1	1

Tableau 7. Métaphores comportementales

1	censure d'information (info)
2	censure d'interlocuteurs (inte)
3	divulgation ex-cathedra (divu)
4	censure thématique (thém)
5	monopolisation d'information (mono)
6	recherche de contacts stratégiques (rech)
7	coalition (coal)
8	désaccord implicite seulement (désa)
9	si déclaration que l'enfant est normal, alors réponse insistante qu'il n'en n'est rien et vice-versa (n.a.)
10	si proposition de services exclusifs, alors insistance pour obtenir encore plus de services (temps 1) ou rejet et insistance pour l'intégration (temps 2) (s.e.)
11	si insistance pour que l'autre parti soit spontané puisqu'on peut l'aider, alors fermeture, affirmation de perte de confiance (s.p.)
12	si accusation qu'un parti est fauteur de trouble, alors l'autre parti déclare qu'il n'y est pour rien, qu'il n'a pas le choix (f.t.)

Tableau 8. Métaphores comportementales

	1	2	3	4	5	6	7	8	9	10	11	12
sy.a	0	1	1	0	0	0	1	1	1	0	0	0
sy.b	0	0	0	0	0	0	0	0	0	0	0	0
am.a	1	1	1	1	1	1	1	0	1	1	1	0
am.b	0	0	0	0	0	0	0	0	0	0	0	0
di.a	1	1	1	1	1	1	1	1	1	1	1	1
di.b	0	0	0	0	0	0	0	0	0	0	0	0
ec.a	1	1	1	1	1	1	1	1	1	0	1	1
ec.b	1	1	0	1	1	1	0	0	0	0	0	0

- GRAPHIQUES:

Les graphiques des figures 1 à 4 ne présentent que les plans les plus significatifs. Pour les métaphores de l'enfant, ce sont les axes 2 et 3 (figure 1) qui illustrent une segmentation nette entre caractéristiques présentes et attendues ainsi que les axes 3 et 4 (figure 2) qui montrent un ordonnancement tant des individus que des caractéristiques (effet Guttman). Pour les métaphores de l'école, ce sont encore les axes 2 et 3 (figure 3) qui permettent d'observer une segmentation nette entre deux rôles attribués à l'école, soit celui d'expert-conseil vs celui de parent-substitut. Enfin, pour les métaphores comportementales, les axes 2 et 3 (figure 4) permettent d'observer la segmentation entre comportements complémentaires et symétriques ainsi que le rôle déterminant d'une caractéristique symétrique (censure d'interlocuteurs) dans l'alternance entre ces deux catégories de comportements.

Figure 1: Métaphores- enfant - 1

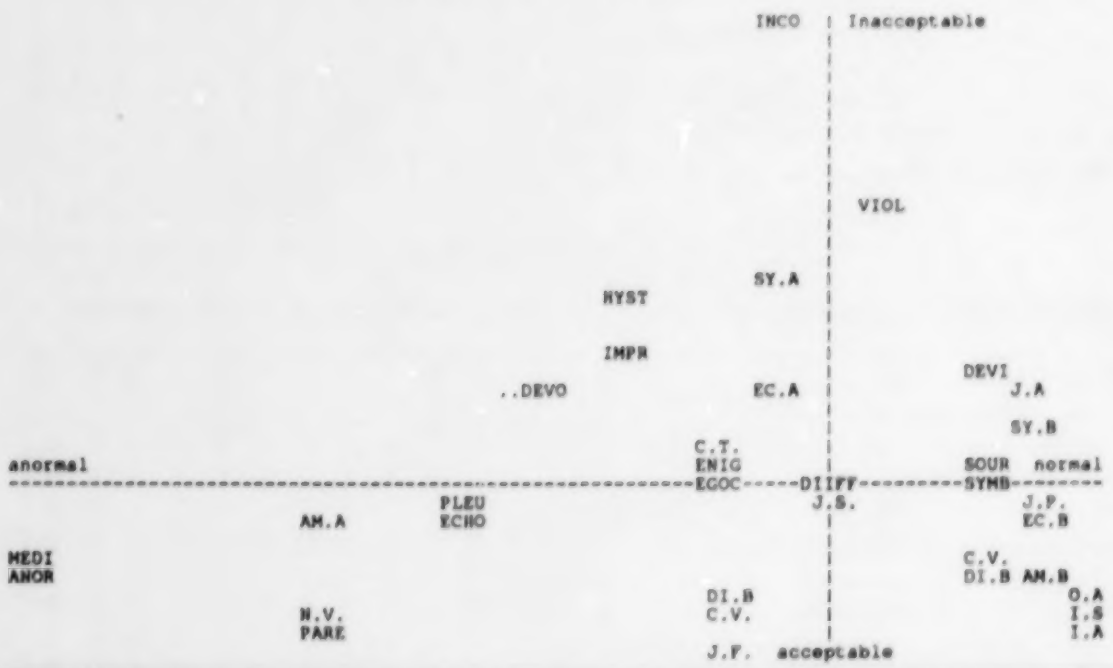


Figure 2: Métaphores- enfant - 2

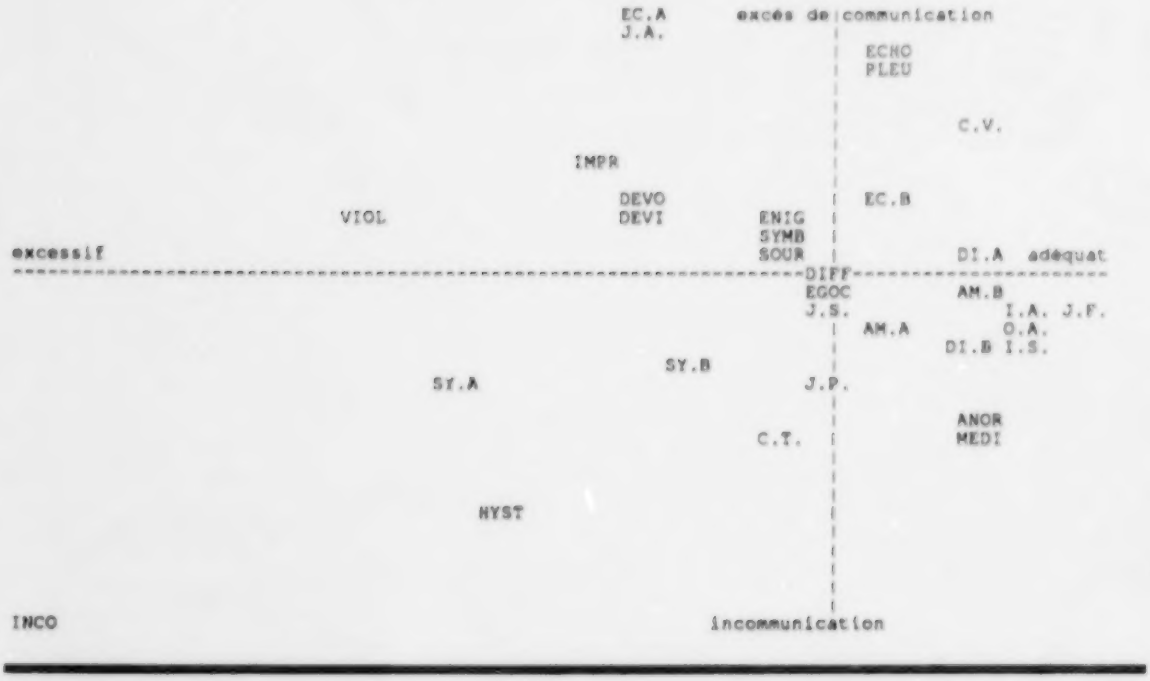


Figure 3: Métaphores- école - 1

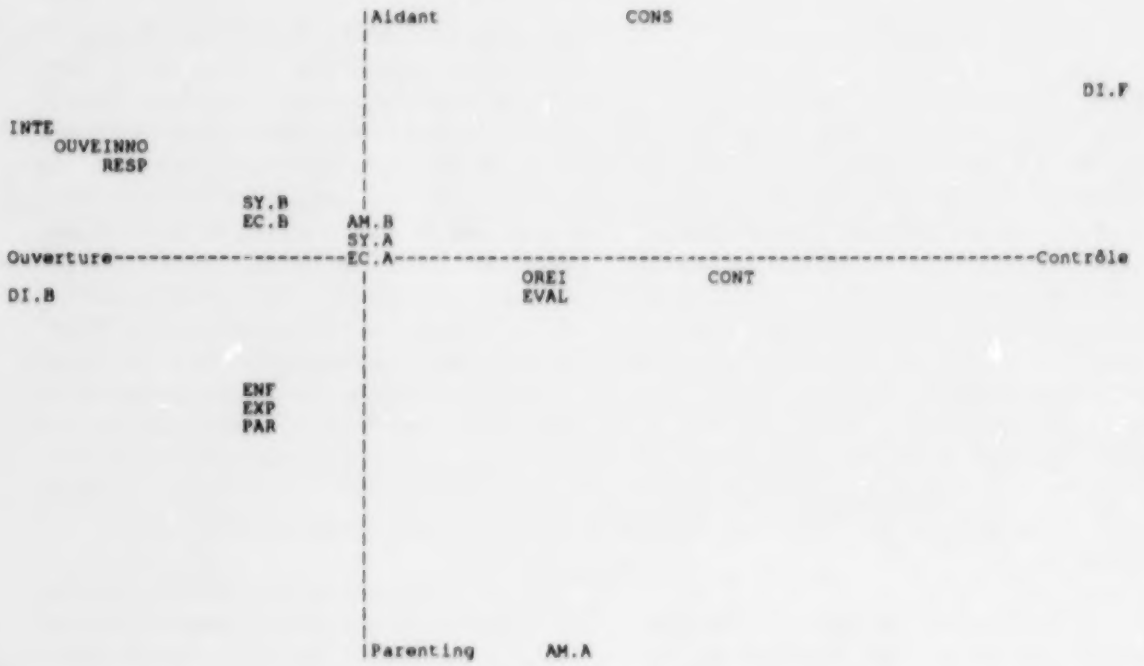
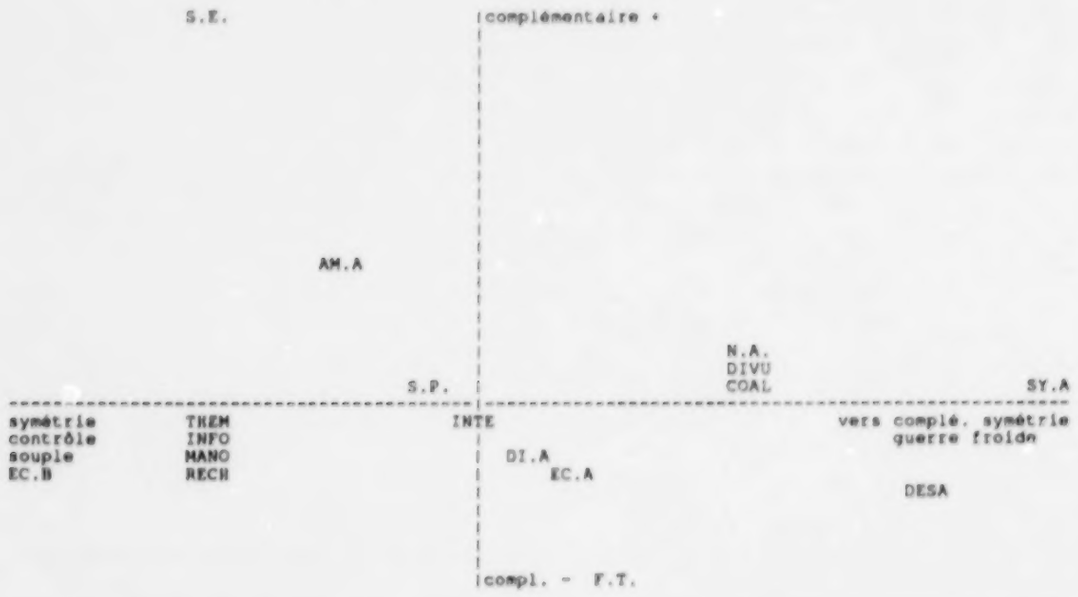


Figure 4: Métaphores- comportements - 1



3.4 Analyse et discussion

- MÉTAPHORES DE L'ENFANT:

La mise en relation caractéristiques - individus illustrée à la figure 1 (axes 2 et 3) démontre clairement que les individus distinguent expressément entre ce que l'enfant *est* et ce qu'il *devrait* devenir. On observe effectivement une segmentation nette entre la gauche de l'axe où des nuages superposés caractéristiques/individus définissent le présent et la droite de l'axe où d'autres nuages superposés définissent le changement attendu. La vision *présente* est constituée de nombreuses caractéristiques typiques de systèmes fermés alors que la vision *attendue* regroupe un grand nombre de caractéristiques d'ouverture.

Or, l'aspect le plus important de ces segmentations ne repose peut-être pas tant sur leurs existences mais sur la relation qu'elles entretiennent avec une famille, la famille ambivalente. Cette famille produit l'écart le plus marqué entre caractéristiques présentes et attendues. Sa projection au présent est d'ailleurs marginale lorsque comparée aux autres ensembles d'individus. Un tel écart et une telle marginalisation me portent à postuler que ce groupe adopte une position utopique, poursuivant un idéal qu'il considère par définition inaccessible. Les fonctions particulières de toute utopie me semblent être de permettre le maintien d'un espoir vital. En faisant du voyage, le but de l'aventure, l'ambivalence comportementale de cette famille prend tout son sens: lorsqu'on ne peut croire à la solution finale, on n'a d'autre choix que d'attendre et d'ériger l'impossible pour pouvoir continuer à être.

Certains s'étonneront peut-être de constater que plusieurs caractéristiques à sens littéral négatif figurent au nombre des caractéristiques attendues alors que certaines caractéristiques à sens

littéral positif comptent parmi les caractéristiques présentes. L'analyse des nuages, mais surtout des points marginaux reproduits, voir figure 2, explique cette apparente aberration.

Ainsi, lors de cette rotation, l'ordonnancement des caractéristiques comme des individus révèle l'existence d'une contrainte bipolaire de sorte que les caractéristiques extrêmes et opposées *(violent, dévorant, déviant, imprévisible* versus *incommunicatif, hystérique, contrôlant)* sont situées aux confins de la courbe alors que les caractéristiques neutres ou à faible charge *(joue avec ses pairs, joue seul, différent, sourire, symbolique, énigmatique, égocentrique)* tombent à proximité du centre de gravité.

On notera que dans l'ordonnancement parallèle des individus, les familles symbiotiques se retrouvent au coeur de cet ordonnancement des caractéristiques. Se dissociant cette fois du groupe scolaire, ces familles semblent envisager que leurs enfants sont capables d'opter simultanément pour des comportements opposés. En faisant appel à la théorie des catastrophes de RENÉ THOM (1977), il est possible de redonner sens à cette vision d'une alternance comportementale apparemment incohérente. En effet, Thom stipule que lorsqu'un organisme éprouve des sentiments contradictoires, une très légère modification dans l'environnement agit comme un point de bifurcation durant lequel deux modes opposés d'action deviennent simultanément possibles. Il est conséquemment probable que ces parents constatent une alternance comportementale chez leurs enfants toutes les fois qu'ils les confrontent avec des attentes irréaliste. En nous obligeant à mettre en relation les ordonnancements des caractéristiques et des individus, l'analyse des données nous force à constater que l'aberrant n'est qu'apparent. Il redevient logique lorsque replacé dans un contexte plus englobant.

Plusieurs caractéristiques neutres, c'est-à-dire celles situées au centre de l'ordonnancement, ne le sont pas dans leur sens littéral; s'y retrouvent des caractéristiques à dénotation positive (par exemple: *joue avec adulte)* comme négative (par exemple: *différent).* Leur position nous oblige à conclure à la présence de sens cachés. Ainsi, l'indifférence générale face à certains termes négatifs — tels *hystérique, contrôle tout, imprévisible, énigmatique...* — me porte à postuler que le droit à la différence exige qu'elle s'opère à l'intérieur d'une latitude tellement restreinte qu'elle est dépourvue de toute émotivité, bref qu'elle se joue à l'intérieur d'un territoire d'indifférence. L'enfant peut différer s'il se limite à être imprévisible, énigmatique, mais non s'il devient violent, incommunicatif, etc....

J'aimerais souligner enfin l'existence d'une association statistiquement significative (qui rend compte de 12% de la variance) entre les deux groupes les plus susceptibles de différer, l'école et les dissidents, quant à leur vision au présent. Une telle association soutenue par une variance quasi nulle quant à leur vision de l'avenir nous oblige à conclure que l'enjeu du conflit n'est pas la réconciliation de points de vue différents mais la guerre pour la guerre ou mieux dit, la guerre pour ne pas voir que l'issue est perdue d'avance. Cette conclusion a beaucoup en commun avec l'utopie des ambivalents. L'étonnante identité des positions entre partenaires officiellement opposés se répètera d'ailleurs dans les deux prochaines analyses. Elle nous obligera à reviser entièrement nos bases théoriques en résolution de conflits.

En résumé, l'analyse des données de l'enfant nous amène à concevoir que *l'enfant est vu comme un système ordonné d'activités qui oscille de façon imprévisible entre des états incompatibles et intolérables de fermeture si bien que tout espoir de résolution définitive paraît improbable et le statu quo inacceptable de sorte que la projection continuelle de la responsabilité est devenue la seule alternative utilisée par les acteurs scolaires et familiaux engagés dans ce projet par définition insoluble pour éviter le désespoir du ghetto de la réalité...*

- MÉTAPHORES DE L'ÉCOLE:

Dans le plan des axes 2 et 3 (c.f. figure 3) les caractéristiques de l'école sont segmentées en deux nuages, celui d'expert-conseil et celui de parent-substitut. Tous les ensembles d'individus reconnaissent et acceptent cette double vocation qui confère à ce milieu une fonction oligarchique. Cette hypothèse est par ailleurs reconnue implicitement dans l'absence de métaphores définitionnelles de la famille dans l'ensemble des corpus. L'école assume seule la responsabilité du changement, à titre d'expert, si bien qu'on n'a plus besoin d'y inclure la famille. Ceci m'amène à postuler l'adhésion implicite par tous les partis au modèle positiviste d'expertise où un groupe remet à l'autre la résolution de son problème.

La prochaine analyse démontre comment on a tenté de part et d'autre de résoudre cette mission impossible...

- MÉTAPHORES COMPORTEMENTALES:

Dans un modèle positiviste, tout système humain menacé cherchera à se protéger en défendant son identité — si *id* signifie chose, alors l'identité, c'est l'entité chosifiée donc immuable — plutôt qu'en la reconstruisant. Les enjeux de maintenance ne sont pas illimités. Ils oscillent entre l'adoption de mesures protectionnistes (coalition, censure) et la création de mesures d'opposition (désaccord systématique, respect de la lettre non de l'esprit).

C'est la position particulière de la famille ambivalente qui permet à nouveau de comprendre la dynamique interactive entre les deux rôles dévolus à l'école: au présent, cette famille voit l'école en tant qu'expert-parent; au futur, elle devra devenir expert-conseil. Si l'on retraduit cette position de la façon suivante — c'est à condition que l'école accepte au départ cette totale prise en charge de l'enfant, que la famille pourra *se représenter* comme parent, représentation entendue au sens double de se présenter à nouveau et de former l'image d'une réalité absente (Le Petit Robert) — on comprend que l'adoption de l'enfant par l'école est la preuve nécessaire que l'enfant est adoptable. Ceci explique l'attitude ambivalente de cette famille mais aussi la position intolérable dans laquelle l'école se trouve placée.

Cette dernière doit réussir l'impossible. Ultimement, elle est totalement piégée: dès que les ambivalents s'associent à elle, l'école ne peut plus accepter sa mission impossible de parent-substitut. Elle ne peut que devenir expert-conseil et viser l'intégration de l'enfant ailleurs; dès qu'elle vise l'intégration, elle détruit sa raison d'être... De là à conclure à l'impuissance de l'omnipuissance...

Pour résumer les résultats de cette deuxième analyse, on pourrait dire que *la classe spéciale de communication est un système d'activités géré par des spécialistes et visant par le développement d'enfants qui n'ont de place nulle part dans la structure sociale, à assurer la réinsertion sociale de leurs familles en construisant, dans un contexte ségrégé, des programmes individualisés de normalisation. Rivée à l'intérieur de sa structure oligarchique, l'école en assumant seule la responsabilité d'une telle opération, se condamne à l'impuissance.*

L'analyse des résultats confirme l'adhésion bipartite école/familles au modèle positiviste. En effet, l'absence de caractéristiques acceptées en mode attendu par les familles indique que celles-ci ne se reconnaissent pas le droit de gérer leur identité puisqu'elles se refusent le droit de s'opposer tant implicitement qu'explicitement. De même, le total synchronisme positionnel de l'école et des dissidents dans toutes les rotations témoignent de leur accord implicite sur les modalités de gestion du conflit. L'analyse illustre donc un différend sans différence, ou même un différend qui ne devrait pas avoir lieu, d'un différend illicite.

L'analyse confirme de plus l'existence distincte des deux stratégies de défense, soit les stratégies symétriques qui forment le premier facteur et les stratégies complémentaires qui forment le second. Leur segmentation, lors de la

rotation des axes 2 et 3, suggère que ces mécanismes s'exercent en alternance. Le passage d'un mode à l'autre semble dépendre d'une caractéristique symétrique — *la censure d'interlocuteurs*. Cette caractéristique fortement corrélée et sise au centre de la projection agit comme un point de bifurcation entre les deux modes. Le durcissement des positions qu'elle implique si on la compare aux autres stratégies symétriques facilite le passage vers des comportements complémentaires où les résistances passent désormais de l'implicite à l'explicite.

Or, ce passage d'un mode à l'autre ne change rien au problçme, revient plutôt à faire encore plus la même chose, la guerre à l'autre qui se refuse de comprendre. Ces comportements permettent de maintenir intactes les règles interactives sans jamais s'interroger sur leurs fondements. Aucune recalibration ne doit apparaître étant donné la nécessité pour les deux partis de maintenir intacte leur épistémè d'expertise qui assure que l'utopie de la cure passe inaperçue.

> *"or, dès qu'une structure se ferme, on peut affirmer que si elle continue d'exister elle ne peut le faire qu'en prenant comme finalité, comme raison d'être peut-on dire, le maintien de sa structure."* LABORIT, 1974: 157.

4 CONCLUSION

Semblables à la pensée psychotique, les métaphores étudiées dans cette recherche illustrent comment un système d'activités humaines qui en vient à ne privilégier que ses structures de contrôle se met à n'entendre que sa voix illuminée. La rigidification de ses structures interactives ressemble effectivement à une hallucination que rend invisibles ses différents niveaux de conscience. En perdant de vue le niveau où il y a symbiose définitionnelle, la tentative de résolution tourne sur elle-même en circuit fermé. Le système aveuglé se met à chercher encore plus son homéostasie sécurisante en s'imposant un modèle du monde complètement prévisible.

Seule une analyse de données pouvait faire ressortir ce type d'enjeu

Références

ALTAN, H. (1979) *Entre le cristal et la fumée*, Seuil.

BATESON, G. (1978) *Mind and Nature, a necessary unity*, Paladin.

BRENT, S. B. (1978) *Prigogine's model of self-organisation in non-equilibrium systems, its relevance for developmental psychology*. Human Development, **21**, 374-387.

LABORIT, H. (1974). *La nouvelle grille*. Folio, Robert Laffont.

MARUYAMA, M. (1983) *Mindscapes and the science theory*. Curre.t Anthropology, **21**, (5), 589-608.

PRIGOGINE, I. et coll. (1972) *Thermodynamics and evolution*. Physics Today (nov.), 22-28.

THOM, R. (1974). *La nouvelle grille*. Folio, Robert Laffont

Des exemplaires du présent document sont en vente à l'adresse suivante:

Centre Interdisciplinaire de Recherches sur
l'Apprentissage et le Développement en Education
Université du Québec à Montréal
C.P. 8888, Succ. A, G-3125
Montréal, Québec
H3C 3P8

Pour commander, remplir le formulaire de réquisition apparaissant à la dernière page du document.

Publié par le Centre Interdisciplinaire de Recherches
sur l'Apprentissage et le Développement en Education
Université du Québec à Montréal
Conception graphique: Pierrette Précourt